Statistical Policy Working Paper 5

Report on Exact and Statistical Matching Techniques

1980
U.S. DEPARTMENT OF COMMERCE
Office of Federal Statistical Policy and Standards

Statistical Policy Working Papers are a series of technical documents prepared under the auspices of the Office of Federal Statistical Policy and Standards. These documents are the product of working groups or task forces, as noted in the Preface to each report.

These Statistical Policy Working Papers are published for the

purpose of encouraging further discussion of the technical issues and to stimulate policy actions which flow from the technical findings and recommendations. Readers of Statistical Policy Working Papers are encouraged to communicate directly with the Office of Federal Statistical Policy and Standards with additional views, suggestions, or technical concerns.

Office of Federal Statistical Policy Standards

Joseph W. Duncan
Director

For sale by the Superintendent of Documents, U.S. Government Printing Office Washington, D.C. 20402

Statistical Policy
Working Paper 5

Report on Exact and Statistical Matching Techniques

Prepared by
Subcommittee on Matching Techniques
Federal Committee on Statistical Methodology

DEPARTMENT OF COMMERCE
UNITED STATES OF AMERICA

U.S. DEPARTMENT OF COMMERCE
Philip M. Klutznick
Courtenay M. Slater, Chief Economist

Office of Federal Statistical Policy and Standards
Joseph W. Duncan, Director

Issued: June 1980

Office of Federal Statistical
Policy and Standards

Joseph W. Duncan, Director

Katherine K. Wallman, Deputy Director, Social Statistics
Gaylord E. Worden, Deputy Director, Economic Statistics
Maria E. Gonzalez, Chairperson, Federal Committee on Statistical

Methodology

Preface

This working paper was prepared by the Subcommittee on Matching Techniques, Federal Committee on Statistical Methodology. The Subcommittee was chaired by Daniel B. Radner, Office of Research and Statistics, Social Security Administration, Department of Health and Human Services. Members of the Subcommittee include Rich Allen, Economics, Statistics, and Cooperatives Service (USDA); Thomas B. Jabine, Energy Information Administration (DOE); and Hans J. Muller, Bureau of the Census (DOC).

The Subcommittee report describes and contrasts exact and statistical matching techniques. Applications of both exact and statistical matches are discussed. The report is intended to be useful to statisticians in various Federal agencies in determining when it is appropriate to use exact matching techniques or when it may be appropriate to use statistical matching techniques. The recommendations of the report also include suggestions for further research.

Members of the Subcommittee on
Matching Techniques

Daniel B. Radner, Chairperson
Office of Research and Statistics, Social Security Administration
Department of Health and Human Services

Rich Allen
Economics, Statistics, and Cooperatives Service
Department of Agriculture

Maria E. Gonzalez (ex officio)*
Chairperson, Federal Committee on Statistical Methodology
Office of Federal Statistical Policy and Standards
Department of Commerce

Thomas B. Jabine*
Energy Information Administration
Department of Energy

Hans J. Muller
Bureau of the Census
Department of Commerce

*Member, Federal Committee on Statistical Methodology

Acknowledgements

The body of this report represents the collective effort of the Subcommittee on Matching Techniques. Although all members of the Subcommittee reviewed and commented on all parts of the report, specific members were responsible for writing different sections. The authors of the respective chapters and appendices appear below:

Chapter	Author(s)
I	Daniel Radner, Thomas Jabine, Rich Allen II
II	Hans Muller, Rich Allen
III	Daniel Radner
IV	Daniel Radner, Thomas Jabine

Appendix	
I	Rich Allen

II Daniel Radner

III Hans Muller, Rich Allen

Maria E. Gonzalez and Thomas B. Jabine provided indispensable guidance and encouragement throughout the Subcommittee's work. Tore Dalenius, an ex officio member of the Subcommittee when the work began, provided important insights in the early stages of the work

and helpful comments on drafts of the report. Others who contributed to the work as members of the Subcommittee in its earlier stages include: Richard Barr, Richard Coulter, David Hirschberg, Matthew Huxley, Benjamin Klugh, Stanley Kulpinski, Robert Penn, and Scott Turner. Members of the Federal Committee on Statistical Methodology and the Office of Federal Statistical Policy and Standards reviewed and commented on drafts of the report. Also, we are grateful to Benjamin Tepping, Ivan Fellegi, Horst Alter, and Michael Colledge for their helpful comments on drafts of the report, and to all those who supplied examples of matching.

Members of the Federal Committee on
Statistical Methodology
(February 1979)

Maria Elena Gonzalez (Chair)
Office of Federal Statistical
Policy and Standards (Commerce)

William E. Kibler
Barbara A. Bailar
Bureau of the Census (Commerce)

Norman D. Beller

Charles D. Jones
Bureau of the Census (Commerce)

Economics, Statistics, and
Cooperatives Service
(Agriculture)

Economics, Statistics, and Cooperatives Service (Agriculture) (Commerce)

Barbara A. Boyes
Bureau of Labor Statistics (Labor)

(Labor)

Edwin J. Coleman
Bureau of Economic Analysis (Commerce)

John E. Cremeans
Bureau of Economic Analysis (Commerce)

Marie D. Eldridge
National Center for Education Statistics (Education)

Daniel H. Garnick
Bureau of Economic Analysis (Commerce)

Thomas B. Jabine
Energy Information Administration (Energy)

Frank de Leeuw
Bureau of Economic Analysis (Commerce)

Alfred D. McKeon
Bureau of Labor Statistics (Labor)

Lincoln E. Moses
Energy Information Administration (Energy)

Monroe G. Sirken
National Center for Health Statistics (HHS)

Wray Smith
Office of the Assistant Secretary for Planning and Evaluation (HHS)

Thomas G. Staples
Social Security Administration (HHS)

Table of Contents

TABLE OF CONTENTS-Continued

Page

CHAPTER IV-FINDINGS AND RECOMMENDATIONS

APPENDICES

TABLE OF CONTENTS-Continued

CHAPTER I

Introduction and Overview

A. Scope of Study

This report discusses matching of data files for research and statistical purposes. Two basic types of matching, exact matching

and statistical matching, are discussed and applications of those two types by various organizations, mostly government agencies, are described. Matching for other purposes, e.g., administrative purposes, is not considered here. In the matching considered here, identification of units, if needed at all, ordinarily is only necessary to make the match. After matching, that identification can be removed. Most of the discussion in this report is in terms of matching records for natural persons. However, similar considerations apply to matching of records for legal persons, for example, corporations, partnerships, fiduciaries. Many aspects of matching for research and statistical purposes have been reviewed by the Subcommittee. Among the aspects discussed in this report are:

- Matching procedures and their development
- Some advantages and disadvantages of alternative procedures
- Confidentiality considerations
- Accuracy of matching results

1. Definitions and Uses of Matching

Although the terms "match," "exact match," and "statistical match"

have been used frequently in the literature, the Subcommittee knows of no generally agreed upon definitions of these terms. For purposes of this report, the Subcommittee has defined a match as a linkage of records from two or more files containing units from the same population. It has defined an exact match as a match in which the linkage of data for the same unit (e.g., person) from the different files is sought; linkages for units that are not the same occur only as a result of error. Exact matching normally requires the use of identifiers, for example, name, address, social security number. The use of the term "exact" match is not meant to suggest that such matches are made without error; problems encountered in carrying out exact matching are discussed in Chapter II. Other terms for exact matching such as "actual" and "object" matching have also been used. The Subcommittee has defined a statistical match as a match in which the linkage of data for the same unit from the different files either is not sought or is sought but finding such linkages is not essential to the procedure. In a statistical match, the linkage of data for similar units rather than for the same unit is acceptable and expected. Statistical matching ordinarily has been used where the files being matched were samples with few or no units in common;

thus, linkage for the same unit was not possible for most units. Statistical matches are made on the basis of similar characteristics, rather than unique identifying information, as in the usual exact match. Other terms have been used for statistical matching, such as "synthetic," "stochastic," "attribute," and "data" matching..1

The definition of a match used here excludes such record linkage techniques as the "hot deck" allocation of values to nonrespondents in surveys because those techniques are considered to involve only one file. Techniques such as matched or paired sampling in experiments are also excluded from the definition..2

Although the definitions used here do not provide a precise dividing line between exact and statistical matching, in practice it is ordinarily clear which matches are exact and which are statistical. From the point of view of accuracy of the matched data, exact matching has ordinarily been preferred to statistical matching. In many cases, for technical or files cannot be carried out. For example, both files

---

.1 The Subcommittee has chosen to use the terms exact match and

statistical match because those terms are the most frequently used, not necessarily because those terms are considered to be the best.

.2 See Althauser and Rubin (1969) for an example of a matched sampling technique.

legal reasons, or both, an exact match between two might be samples which have few units in common. Legal restrictions on exact matching, which have existed for some time, have been increasing in recent years (e.g., the Privacy Act of 1974 and the Tax Reform Act of 1976). These limitations on the use of exact matching have led to further interest in alternative methods of matching. In practice, the choice between exact and statistical matching sometimes is a choice between statistically matching easily obtainable files which cannot be exactly matched and exactly matching files which are not as easily obtained (especially with identifiers). In some cases files which can be exactly matched are obtainable but contain data which are less appropriate for performing the desired statistical analyses. The impetus for the formation of this Subcommittee came from

restrictions on the use of exact matching arising from confidentiality considerations. The original question to be examined was to what extent and under what conditions is statistical matching an acceptable alternative to exact matching. Thus, the Subcommittee did not examine alternatives to exact matching other than statistical matching. Although a comprehensive comparison between exact and statistical matching was originally intended, the Subcommittee determined that such a comparison was not possible at this time because so little is known about the error structure of statistical matching procedures. For this reason, the Subcommittee decided to summarize in this report what is known about exact and statistical matching, to give examples of applications of both types of matching, to make some limited comparisons of exact and statistical matching, and to suggest directions for future research.

2. Matching Applications and Examples

Matching of data files for research or statistical purposes ordinarily is a step in the preparation of the data needed to perform statistical analyses. In assessing the data needed for a given analysis, there often are cases in which one existing data set does

not contain all of the variables needed (or contains variables of less than sufficient accuracy). Several different approaches can be used to deal with this problem. One possibility is direct data collection of all the needed variables, for example, in a sample survey. Another possibility is the assignment or imputation of values using statistical techniques such as regression analysis (perhaps using information from another data file). A third possibility is matching two or more existing data sets to add the desired variables, using either exact or statistical matching. Thus, matching is merely one of a larger group of techniques which can be used to add variables needed to perform statistical analyses. However, there may be cases in which matching, specifically exact matching, is the only feasible method of preparing the needed data. For example, cumulative health histories of sufficient accuracy might require the exact matching of hospital records. here are also cases in which a comparison of the presence of units in two files, rather than the addition of variables, is needed. In this type of application, there are few, if any alternatives to exact matching. Where the goal is the construction of a multipurpose file, rather than performing a specific analysis, exact and statistical matching

can be particularly appropriate because large numbers of variables can be added relatively easily using matching. The Subcommittee collected many examples of matching of data files. As noted above, the applications can be divided into two broad categories: (1) adding to a base file more variables or additional reports on the same variables; and (2) comparing the presence of units in two files. Within type (1) several different kinds of applications can be identified. One application is the addition of more variables to enrich analyses or to make possible analyses which otherwise could not be done. Both exact and statistical matching have been used in this application. A cross-section example of one such exact match is the addition of Social Security Administration (SSA) age, race, and sex data to Federal individual income tax return records in order to make it possible to analyze income and tax data by those characteristics. In another cross-section example, a statistical match was carried out between observations from a household survey and a sample of Federal individual income tax returns in order to add more detailed and more accurate income information to the household survey data (Budd, Radner and Hinrichs, 1973). A longitudinal example of exact matching is the linkage of hospital admission and separation records into cumulative health histories (Smith and

Newcombe, 1975). Another kind of application within type (1) is the evaluation of data, in which initial variables are compared with added variables, or with additional reports on the same variables- from other existing sources or from special evaluation surveys. Evaluation of the accuracy of data was carried out using the 1973 Current Population Survey-Internal Revenue ServiceSSA Exact Match Study. In that project, the income data from the different data sources were compared

and response and reporting errors were analyzed (e.g., Alvey and Cobleigh, 1975). Definitional differences were examined in Sweden using exact matching. Two different definitions of unemploymentfrom a household survey and from the labor market board-were compared by matching survey responses and labor market board records.

In type (2) (comparing the presence of units in two files), two different kinds of applications can be identified: evaluation of coverage and construction of more comprehensive lists. The Bureau of the Census has conducted numerous coverage evaluation studies in connection with the Decennial Censuses. For example, in connection with the 1960 Population Census, samples from 1950 Census records, registered births, and other sources were matched with 1960 Census records, and coverage was assessed (Perkins and Jones, 1965). In such matches, the emphasis is upon the presence of units in the files, rather than upon the relationships between data in the two files. In an example of list construction, the Economics, Statistics, and Cooperatives Service (ESCS) of the U.S. Department of Agriculture uses exact matching in the construction of a master list sampling frame of farms in each state. This master list was constructed from several different lists, and exact matching was used to detect duplication between (and within) the different lists (Coulter, 1977). Statistical matching is not appropriate for type (2) applications.

In most of the applications mentioned above, one possible effect of matching was a reduction of response "burden". That is, to

collect the same information without matching would have required a considerable amount of direct data collection. Also, in some of those applications, cost reduction was a beneficial effect-i.e., matching was less expensive than direct collection of the same combination of data would have been. The Office of Federal Statistical Policy and Standards (1978a) suggested the use of statistical matching to reduce response burden and cost by means of what are called "nested surveys." In such surveys, different samples from the same population are asked separate sets of questions, with a core of questions in common. The data from these different samples can then be statistically matched to obtain relationships among the items not in the common core of questions.

3. Confidentiality Issues

As noted earlier, legal restrictions on exact matching have led to increased interest in alternatives to exact matching. The relevant confidentiality issues are discussed in this section. Exact matching of records for individual reporting units for statistical research purposes raises two important questions in the area of

confidentiality:

To what extent should such matching activities be conditional on the "informed consent" of the individuals whose records are being matched?

To illustrate this issue, consider the case of a statistical survey in which participation is voluntary and information is to be collected on topics such as income, assets, use of medical services, voting behavior, etc. To measure the validity of the survey responses, they will be individually matched to and compared with relevant information in administrative record systems of tax collection agencies, banks, hospitals, and others.

Such record checks (including reverse record checks, where the sample of persons to be interviewed is drawn from the relevant administrative system) have been a valuable tool for the improvement of survey methods. Full respondent knowledge of the nature of the study and the procedures to be followed might condition their responses and to some extent defeat the purpose

of the study. Nevertheless, both ethical and legal considerations require that individuals providing data be adequately informed of the uses that will be made of the data they provide.

Do the benefits to be gained by exact matching outweigh the risks inherent in assembling large amounts of information about individuals in a single location?

When large amounts of information about an identifiable individual are available in a single file, the potential for use of the information to the detriment of that individual is greater than if the information were segmented and the parts maintained in different locations. Some exact matching activities conducted for statistical purposes have brought together large amounts of information for identified individuals, from both survey and administrative record sources.

Although the creation of such files clearly increases the potential for harm to individuals, it is also relevant to ask

whether any individuals have, in fact, been harmed as the result of disclosures from matched data files created for statistical purposes. Inquiries made by another group (Office of Federal Statistical Policy and Standards, 1978b) have not identified any such cases.

These and related concerns have led to the creation of an environment in which significant restrictions have been placed on the exact matching of records belonging to more than one Federal agency and on the matching of Federal agency records with those of other organizations.

The Privacy Act of 1974 placed certain limitations on the disclosure of individually identifiable records in the hands of Federal agencies. In brief, these limitations have the following

effects on exact matching for statistical purposes:

. Identifiable records can be disclosed (transferred) within an agency on a need to know basis. For purposes of the Privacy Act, each Department (e.g., HHS), is an agency, so that intra-departmental matches can be carried out if not otherwise prohibited by law.

. Identifiable records can be disclosed to the Census Bureau for use in its census and survey activities. Subsequent to the Privacy Act, revised Census legislation placed reimbursable work conducted by the Census Bureau for other agencies in the category of Census activities to which this provision applies.

. Identifiable records can be disclosed to any agency or organization under a routine use established for that system of records. The routine use is established by the agency controlling the source record system, and the use for which the disclosure is to be made must be deemed "compatible with the purposes for which it was collected". There may be problems in exercising the routine use provision where the planned match requires the exchange of identifiable records in both directions (Jabine, 1976, p. 229).

In addition to the general restrictions imposed by the Privacy Act, there are several agency statutes which further limit the ability to conduct interagency matching studies. Some statistical agencies, in particular the Census Bureau and the National Center for Health Statistics, have statutes which prohibit the transfer of identifiable records to any other agency or organization. The Tax Reform Act of 1976 limits the release of tax return information, broadly defined, for identifiable individuals and legal persons to certain agencies, uses and types of information specified in the law. One example of the effects of these new restrictions is that most researchers conducting follow-up studies no longer have access to IRS records to determine which members of their study populations are still alive and where they are located. Consideration of the issues and problems described in this section has led many persons to advocate greater use of alternatives to exact matching to achieve desired ends, or at least to examine the feasibility of alternative methods. Statistical matching has been used in some situations where exact matching was not feasible; the question has been raised in some quarters as to whether it should be used even where exact matching is feasible. For example, Duncan (1976) recommended that consideration

be given to the use of statistical matching and to research on the merging of grouped data to t0 estimate the relationships among variables without matching individual records.

4. The Role of Computers

Modern computers and development of advanced software for matching have made many matching applications feasible which could not be done manually. Exact matching has been performed manually and by computer. Exact matching by computer, once the source materials are in machine readable format, is much faster and less expensive than performing the same matching manually, but the biggest advantages arise from consistency of decisionmaking and use of more complex matching rules. For example, in a manual match of name and address files, ordinarily last names are reviewed, then first names of individuals with the same last names, then addresses, etc. A computer match procedure can compare all elements in one pass, assigning agreement and disagreement weights to each element. Some matching examples in this report involve comparison of 15 or more variables which would not have been feasible by manual procedures.

There do remain some situations in which manual matching is more practical or possibly more successful than computer matching. In Chapter 11, D, under Practical Problems, there is some discussion of a few of these situations. Statistical matching has only been performed by computer; it would not be practical to carry out statistical matching manually.

B. Auspices

This report represents the collective effort of the Subcommittee on Matching Techniques of the Federal Committee on Statistical Methodology, which operated under the auspices of the Office of Federal Statistical Policy and Standards, Department of Commerce (previously the Statistical Policy Division,

Office of Management and Budget). The group was formed in early 1976 as one of two working groups of a Subcommittee on Confidentiality Issues chaired by Thomas B. Jabine. The working groups were subsequently given separate subcommittee status. The other group, the Subcommittee on DisclosureAvoidance Techniques, issued its report in May 1978 (Office of Federal Statistical Policy and Standards, 1978b). The opinions expressed here reflect the collective judgment of the Subcommittee and do not necessarily reflect those of the Federal Committee on Statistical Methodology or the Office of Federal Statistical Policy and Standards.

C. Dissemination of Report

This report is intended for circulation to agencies and Federal offices which may utilize matching techniques. However, a broader audience may be interested in the report. The report attempts to present the major considerations and concerns for the use of matching procedures. Examples of present and past applications are included to aid the reader in visualizing the types of files which can be

linked and the types of variables needed for matching.

D. Organization of Report

Chapter II contains a discussion of exact matching. That discussion includes a brief overview of the nature and history of exact matching, a description of the steps in exact matching procedures, and descriptions of practical problems and reliability. A detailed example of exact matching is presented in Appendix I and summaries of selected examples are shown in Appendix III. A discussion of statistical matching is presented in Chapter III. Because statistical matching is not a very well-known technique, in Chapter III substantial space is devoted to the nature of statistical matching, and summaries of many statistical matches are included. Discussions of criticisms of statistical matching and types of errors in statistically matched data are also presented, although those discussions are necessarily sketchy since little is known about the reliability of statistical matching. Appendix II contains a detailed example of statistical matching. Chapter IV contains the findings and recommendations of the Subcommittee. The findings are concerned with definitions, usefulness, and applications of matching, as well as

errors in matching and confidentiality considerations. The general recommendations involve the use of matching, documentation of matches, public release of matched data, and confidentiality restrictions on matching. Also, further research on both exact and statistical matching is recommended. A bibliography of exact and statistical matching references is included at the end of this report.

CHAPTER II

Exact Matching

A. Nature and History.3

As defined earlier, an exact match is a match in which the linkage of data for the same unit is sought. Exact matching ordinarily is

carried out using a set of characteristics ("identifiers") contained in both records. The unit may be a person, family, housing unit, address, farm, business firm, and so fortb, or it may be an event such as a birth. The following observations refer mostly to person matching but they could be applied or adapted to other units as well. Usually, the records come from two different sources (files). Three or more files may be involved, but even in that case the matching is often carried out between two files at a time; however, procedures have been developed for matching multiple files simultaneously to end up with a single unduplicated file (see Appendix I of this report). In some cases, all units (and no others) are assumed to be represented in both files; in others, one file may represent a subset of the other one; or the two files may overlap but may each include a number of units not covered by the other. In the following, matching is described in terms of linking records from a "base file" to those in a "reference file". Matching in both directions may be indicated in some circumstances; the procedures for two-way matching are a simple extension of those for one-way matching. (When one file is a subset of the other, exact matching is feasible only from the subset to the complete file.) "Exact matching" is not necessarily

"exact" in the sense that there must be exact agreement on all characteristics that are compared. The source files usually include some incomplete records and some inaccurate data. Allowances must be made for this at various stages of the matching process. Exact matching techniques therefore are not just procedures for bringing together two records that are clearly and uniquely identified and unequivocally known to refer to the same unit. Exact matching can be practically error free under favorable conditions (for instance, when matching two files on the basis of social security numbers that were transcribed from reliable records rather than reported from memory); but under less favorable conditions some uncertainty about the results of the matching must be expected, that is, the matches obtained will probably include some erroneous ones, and some true matches will be missed. The matching procedures should be designed to control matching error in such a way that the error in the conclusions to be drawn from the study will be kept at a tolerable level. Thus the procedures must be adapted to the conditions prevailing in each project, with respect to the objectives of the study and the quality of the source files (and, as always, the human, technical, and financial resources and, in some cases, time

constraints). In general, with more incomplete and inaccurate source files, more complex matching procedures are called for and a higher proportion of matching errors may be unavoidable. Exact matching, in its simplest form, has been known for many years. For example, for quite some time there has been interest in matching a list of current taxpayers against the previous payee list or a list of units which should be paying taxes. However, in the context of this report this type of example normally is not for statistical purposes and is excluded from consideration. Some of the earliest applications of exact matching techniques for statistical purposes have been for follow-up studies of Census data. Appendix III, Reference A describes the procedures used to match 1960 Population Census Records against 1950 Population Census Records, Registered Birth Records, 1950 Population Evaluation Survey results, and Alien Registration Records. This match involved a clerical reverse record match procedure on addresses. Codes were given to the various name, address and supple

---

.3 Marks et al., 1974; Steinberg and Pritzker, 1967.

mental information items to characterize the amount of agreement. Each comparison case was then considered as matched or nonmatched. The simplest clerical matching techniques utilize comparisons of names only. The development of computer capabilities gave rise to exact matches on identifiers rather than names. In the United States social security number (SSN) has been extensively used for exact matches of separate files. Several of the examples in Appendix III used only SSN for matching. A number of individuals have conducted research in theory and procedures for exact matching of files. The paper by Fellegi and Sunter (1969) expressed a record linkage theory involving probabilities for the matched and unmatched sets of units from two files. The Economics, Statistics, and Cooperatives Service, USDA, exact match example in Appendix I bases much of the linkage techniques on FellegiSunter. Similar techniques were also used for the Statistics Canada applications included in Appendix III, references L and M.

B. Types of Matching Error.4

In practice it is almost inevitable in most matching projects that some matching errors occur, even with the most sophisticated procedure and the most careful execution. These errors fall into two major classes:

a. Erroneous match ("false match", "positive error", "Type II error"): Linking of records that correspond to different units.

b. Erroneous non-match ("false non-match", "negative error", "Type I error") : Failure to link records that do correspond to the same unit. "Gross matching error" is the sum of both types of error.

"Net matching error" is their difference. However, this concept is useful only in certain applications, mainly in coverage evaluation, where the objective is the estimation of the true size of

a population. When the goal of the study is the estimation of other population parameters, the "net error due to matching" may be a more complex function of the two types of error, depending on bow each type affects the estimates.

Erroneous matches may be of two kinds:

a. The reference file includes a true match for a certain base record but the latter is mistakenly linked not to its true match but to a different reference record.

b. The reference file does not include any true match for a certain base record but the latter is mistakenly linked to some reference record.

The term "mismatch" is used by some for any erroneous match, by others in a more restricted sense for the (a) kind only. While the (b) kind of erroneous match is always unacceptable, the (a) kind may be considered as acceptable matches in some studies but not in others, depending on the objectives of the study. For example, in one-way matching, a base file unit for which there is a true but

undetected match in the reference file may be classified as "matched" on the basis of an erroneous linkage with the reference file record of a different unit (a "mismatch" in the strict sense of the(a) kind). In a coverage study in which the only objective is to determine whether each base file unit is present in the reference file or not, that mismatch would be acceptable. The same mismatch would be unacceptable, however, when the objective is the comparison of certain characteristics reported for the same unit in the two files or the addition of data from the reference file to the matching record in the base file. The relative importance of each type of error varies depending on the objectives of different projects. Content evaluation and other studies based on comparisons of characteristics of matched pairs require a low Type 11 error, that is, high confidence in 'matched" pairs being true matches; Type I error (failure to find some true matches) will not affect the findings derived from the matched pairs unless the characteristics under study are distributed differently in the matched and the erroneously not matched records. In coverage evaluation, on the other band, both types of error affect the results-in opposite directions- and the desired procedure is one that leads to a balance between both types of error, resulting in a tolerably small net error. (However,

if Type I and II errors were both very large the procedure would be suspect, even if it resulted in a very small net error.) The foregoing considerations must be kept in mind when choosing the match procedures for a particular project. The ways in which the procedures can be adjusted to serve the purpose of each study are treated in Section C of this chapter.

.8 Marks et al., 1974; Seltzer and Adlakha, 1969.

C. Procedures.5

In general, exact matching requires the following steps:

1. Preliminary steps: Improvement of the quality of source files; elimination of outof-scope records; standardization of files.
2. Selection of match characteristics (components), and definition of "agreement" and "disagreement" (tolerance limits) for each characteristic.
3. Blocking (comparison reduction) and searching (identification of comparison pairs).
4. Weighting of characteristics of comparison pairs.
5. Determination of thresholds for designating "matches" and "non-matches" (or three groups: match, non-match, undetermined).
6. Validation of decisions; follow-up on undetermined cases (reconciliation).

In practice, these may not always be recognizable as distinct steps, but explicitly or implicitly, they are usually carried out in some form. The procedure must be designed for each project, on the basis of previous experience with the same or similar source files, or of a special pilot study, or of early data from the study itself (in which case tentative match rules must be set up initially based

on whatever information is available at the outset). The decisions needed at each step may be taken on an intuitive, empirical, or mathematical basis. "Intuitive" decisions are based on the researcher's experience with or knowledge about the same kind of files and his best judgment of the quality and discriminating power of the data. "Empirical" decisions are derived more formally from actual matching results from similar studies or, preferably, directly from the study itself, either through a pilot study or a sample of the main study. "Mathematical" decisions are derived from mathematical models of the matching procedure in the given set of files, using prior knowledge or assumptions about the probability of occurrence of various observed data configurations in true matches and true nonmatches. The more complex procedures are not necessarily always the best ones; the choice must be made in terms of the source data, the objective of the study, the precision required in the output, the resources available, cost and time limitations, etc. The nature of the project is also a factor: in a continuous or multiround project the initial period can be used for testing and improving the match rules; for a onetime project of short duration a pilot study is essential, or else, if the main study is small, it might be carried

out like a pilot study, with very thorough follow-up so that the effect of different matching rules can be investigated. The entire procedure for a particular study should be oriented towards the goal of minimizing (or reducing to a tolerable magnitude) the error in the conclusions of the study.

1. Preliminary Steps

In many cases the researchers have no control over the quality of the source files. However, where one or both files are collected especially for the matching project, the results of the matching can be greatly . proved by intervening in the forms design, training ofmf interviewers, and so forth, to make sure that characteristics that will facilitate the matching are included, and that the interviewers understand the importance of complete and accurate information for those characteristics. Elimination of out-of-scope records may be necessary in some cases, if the source files do not cover exactly the same area or time period or population group. Examples: uncertain area boundaries; inclusion or exclusion of institutional population or Armed Forces; and so forth. Out-of-scope records in one file cannot possibly be matched in the other file and should be eliminated

at the earliest possible stage, to keep them from being counted as nonmatches. Standardization of the files is not as critical in clerical matching as in matching by computer. To be matchable by computer, one or both files may have to be reformatted.

2. Selection of Match Characteristics (Components), and Definition of "Agreement" and "Disagreement" (Tolerance Limits) for Each Characteristic.6

In many match projects so little information is available for matching that all of it must be used in the matching process. In others there may be some redundant information, and the "best" characteristics can be chosen as a basis for the matching decisions. The selection should be based on the quality of the available data, the discriminating power of the various characteristics, and the purpose of the study. Ideally, the most accurately reported and the most

.5 Marks et al., 1974; Appendix I of this report.

.6 Madigan and Wells, 1976; Housni et al., 1978; Nathan, 1978; U.S. Dept. of Commerce, 1977.

discriminating characteristics would be preferred, but there may be a conflict between these two requirements. (Social security numbers actually assigned are close to being a unique identifier = 100% discrimination; however, social security numbers obtained in household surveys contain a sizeable proportion of errors.) The less discriminating power a characteristic has, the less information it provides, and the more characteristics must be compared before a decision (match or nonmatch) can be made. Because reporting in the source files is not always accurate, insistence on exact agreement

between two records would lead to erroneous nonmatches. The match rules should allow some tolerance, such as age differences of plus or minus one or two years, common spelling differences in names, etc. On the other hand, if the tolerances are too wide, erroneous matches will result. The selection of the match characteristics and the setting of tolerance limits for each characteristic should be done so as to minimize the type of error that should be kept low in order to best serve the purpose of each project. Various more or less elaborate procedures for doing this have been described in the literature; they may be based on the researcher's past experience and judgment, or on thorough analysis of a pilot study or a sample of data from the project itself; such an analysis would require a more thorough investigation of potentially matched records than is generally possible for an entire project, in order to establish the characteristics of true matches (and nonmatches) with a high degree of confidence. Operational efficiency should be considered also; if there is a choice between several characteristics or tolerance limits that are about equally efficient in terms of keeping the critical type of matching error low, the selection should be made in terms of operating considerations, such as cost, difficulty, and risk of error

in the implementation.

3. Blocking and Searching.7

Searching in the reference file for a record or records that might match the input record can be viewed as reducing the possible comparison pairs (each input record paired with all reference records, one at a time) to a number of comparison classes, each class having some common characteristics and including a more manageable number of comparison pairs that will then be compared on their other characteristics. In matching by computer, this Is important to keep the cost down; it is achieved by "blocking" the files through the use of Soundex or similar code systems for names, or of geographic codes (street segments, enumeration districts), and so forth, with the effect that each input record will be compared in detail with relatively few reference records. However, the saving must be weighted against the risk of increasing the number of erroneous nonmatches: a reference record that agrees with an input record on all characteristics except the one used for blocking may in fact be the true match for the unit record, but because it is not included in the right block it will not be compared with the right unit record

and both records may be classified as not matched (or they may wind up being paired with the wrong partners). This can be avoided to some extent by multiple matching: the records not matched according to one set of criteria are processed again using a different set. Obviously, that would increase the cost. In manual matching, blocking may not be a separate step but is implicit in the search operation. For example, in matching by name, the clerk will use only that part of the reference file that includes the names starting with the same letters as the input record, and so forth. In general, the larger the blocking unit, the higher the cost of matching within blocks and the greater the risk of erroneous matches; the smaller the blocking unit, the lower the cost of matching within blocks but the greater the risk of erroneous nonmatches. Ideally, blocking should be done on the basis of characteristics which will virtually never disagree in the case of true matches; they should also disagree nearly always in the case of nonmatches. The combination of two characteristics may be most effective, e.g., father's name and mother's maiden name (double Soundex code). The characteristic used for blocking should preferably be independent of the other matching characteristics (e.g., blocking by geographic characteristic, matching by name, etc.); if it is not

independent (e.g., blocking by Soundex, matching by full surname), this fact must be taken into account in defining the matching rules.

4. Weighting of Characteristics of Comparison Pairs.8

After blocking, the characteristics of the input record are compared with those of the reference

---

.7 U.S. Dept. of Agriculture, 1977; U.S. Depart of Commerce, 1977

.8 Perkins and Jones, 1966; Smith and Newcombe, 1975; Fellegi and Sunter, 1969; Tepping, 1968; USDA technical papers cited in Appendix I of this report.

records in the corresponding comparison class, and the "best match"

is selected from those records. Whenever more than one characteristic is compared, the fact that the various characteristics contribute different amounts of information must be taken into account. For example, for deciding whether the two records of a comparison pair refer to the same person, agreement on sex contributes less information than agreement on names; among names, agreement on a common name contributes less than agreement on an unusual name. These differences can be taken into account through a system of weighting. Weights can also reflect the amounts of information derived from different degrees of agreement on one characteristic, such as exact agreement on year of birth or a difference of plus or minus I year, 2 years, and so forth. As a general rule, more weight is given to items with high discriminating power and low error rates. The weights can be derived from a set of explicit and detailed rules, or they can be based on the judgment of the person doing the matching as to the relative importance of the observed kind and degree of agreement in each comparison pair. Explicit rules, in turn, can be formulated intuitively or they can be derived from a mathematical model of the matching process; in either case, some knowledge about the behavior of the matching

characteristics is needed, either from previous studies with similar data, or from a pilot study, or it may be derived in the course of the processing from the data under study. It should be noted that, for some characteristics, agreement and disagreement do not carry equal weight (in opposite directions). For instance, agreement on sex is not very conclusive evidence of a match, but disagreement on sex is rather strong evidence against a match. Disagreement as well as agreement can be included in the weighting system; negative weights are assigned as evidence against a match. For each comparison pair, the weights assigned to the various match characteristics are combined into an overall score in order to select the "best match" among the pairs in each comparison class (block). In classes with only one comparison pair there is no choice, but the match data may need to be weighted in any case for the following step.

5. Determination of Thresholds.9

The "best match" among the pairs in a comparison class (or the only pair in a class) is not necessarily an acceptable match. It is accepted as a match only if its level of agreement is higher than a designated "threshold" level. As with other matching decisions, the

threshold can be defined intuitively on the basis of previous experience and knowledge of the data sets involved, or it can be derived formally from a mathematical model. The important criterion is that this step, in conjunction with the other parts of the matching procedure, should lead to the goal stated before, that is, to minimize (or keep tolerably low) in each study the error of estimation of the population parameters that are of interest in that study. Ultimately, all comparison pairs should be designated as "matched" or "unmatched", making sure that no reference record is matched to more than one record. If some follow-up is feasible, the final decision may be improved by initially defining two thresholds- an upper one above which a pair is considered as matched, and a lower one below which a pair is considered as not matched. The pairs falling between the two thresholds can then be followed up either by a thorough re-evaluation of the available information by an experienced researcher, or by repeating the matching process but including additional variables available in the records, or by additional field work to reconcile conflicting information in the records or to obtain additional information. In any case the follow-up work should lead to a final decision of "matched" or "unmatched".

6. Validation of Decisions

If the source files were perfect-with complete and error-free identifying information-matching problems would be controllable. As it is, the results will usually be affected by the previously described uncertainties implicit in matching with imperfect data. As a general rule, a matching project should include a validation of the matching decisions and an evaluation of the remaining matching error. This could take the form of an intensive study, including field follow-up if at all possible, of a sample of "matched" and "unmatched" records, endeavoring to ascertain their true status. If pilot studies were undertaken at earlier stages (for decisions on matching characteristics, tolerances, weights, thresholds) , their results may be useful for this purpose also and may reduce, if not eliminate, the need for more field work. The findings from the sample or pilot study-as to the proportion of each original match status group that were found to be true matches or nonmatchescan then be used to estimate the matching error remaining in the entire file.

_____

.9 References: see C. 4.

.10 Scheuren and Oh, 1976; Seltzer and Adlakha, 1969.

If the evaluation indicates that certain match status the probability that the matched records refer to the groups have a very low error rate and certain others same unit is very high. There is less certainty about have a high one, and if an extensive follow-up is feasible (by mail, phone, personal interview, or record search), a full follow-up may be undertaken only for the group with the high error rate, in order to obtain more information that may either confirm or change the match status and give the validated status a higher probability of being correct. At least a sample of the other status groups should be followed up the same way, to avoid the possibility of bias arising from special treatment for one group.

More sophisticated methods of estimating the matching error have been devised. When the matching procedure is based on a mathematical model the estimation of the error probabilities is an integral part of the procedure. With some models the admissible error rates for each match status group may be specified to begin with and the match rules chosen to give results with the specified error rates. Given the probability that some "matched" records really refer to different units and that some "unmatched" records really have a match in the other file, the conclusions drawn from the results of the matching are also subject to error because of these matching errors. (They may also be affected by other error sources, such as different concepts used in the source files for a variable that is to be compared between the two files, or coverage differences between the files.) Attempts can be made to adjust the results, on the basis of prior knowledge or assumptions about the true distribution of some characteristics. Such adjustments have been designed specifically for some studies.

D. Practical Problems

1. Source Data

In practice, most if not all match projects are affected in some degree by imperfections in the source files-outright errors in the data; spelling variations; absence of some data from one file or the other; differences in concept between apparently comparable data; variability in data reported by different respondents, at different times, or for different purposes; inclusion of units that should not be included and omission of units that should be included. Recent legislation has restricted the use of the best identifiers (names, social security numbers) in some cases. Generally, if a match is based on a sufficiently discriminating combination of several characteristics, failure to match: it could be due to an error in either file or to a true change in some match characteristic if the source files refer to different dates. One wrong digit in an identification number, or in a house number if the first search must be based on the address, can cause an erroneous classification as "nonmatch"; so can a misunderstood or misspelled name (unless it is one of the common spelling variations that are taken into account in the name coding schemes), or a change of address or (for women) a name change due to marriage or divorce. In some studies, the problem

of changing data can be reduced to a reporting problem by asking for previous addresses and previous names (maiden name, former married name) when the data for the later file are collected.

2. Matching Procedures

Problems can arise if the purpose of the study is not kept in mind at all stages when the matching procedure is designed. A procedure that is best for one study may distort the conclusions from another study that has different objectives. The execution of the procedure is beset with other kinds of problems. Except when the matching decision can be based on a simple and practically unique characteristic, such as a well-reported identification number, the matching rules are bound to be complicated.

3. Matching mode (manual or computer)

A computer program for matching requires very detailed rules for tolerances, weights, etc., which is normally an advantage in that the matching decisions will be uniform, not subject to different

interpretation by different clerks. It may be a disadvantage if there is supplementary information in the records that does not lend itself to coding or could not be included in the computer program for other reasons, but could be used by an experienced person to decide for or against a match when the basic information is ambiguous. For instance, sometimes the question whether two records refer to the same person may not have a clear answer if only the information in the two records is compared; but if the records are part of household or family groups the information about household composition (relationships, birth order, etc.) and about the other household members may provide the answer. These intrahousehold relationships can take so many different forms that they could not possibly all be included in a computer program. Similarly, an experienced reviewer will

often detect some misspellings that would escape matching by even the most sophisticated name coding routines.

The advantage of the greater speed of a computer for matching may be lost if the records are not computerized to begin with and require a large amount of manual preparation (coding, keying, etc.) to make them machine readable. Certain items (especially addresses) may also need reformatting in one or more files before they can be compared by computer; that would require additional programming and computer time. In some applications manual matching may be less costly. For example, the determination if 2000 individuals are included in a nationwide, well-indexed file of many millions of records will be cheaper by manual look-up than by processing the entire file by computer (unless the matching can be done while the large file is passed through the computer anyway for some other purpose). In some cases it may be possible to take advantage of the best features of both computer and manual modes by doing the work in two stages:

1. Computer match of the entire file, using criteria that will identify matches and nonmatches with near certainty,

leaving a portion of the input file unclassified (if the identifying information is reasonably good, this should be a small proportion).

2. Manual review of the unclassified portion, making use of any available information not included in the computer program, possibly using additional files that are not machine readable.

4. Follow-up

Like the matching procedure, the follow-up procedure must also be designed to fit the purpose of the study. In addition, it must fit the matching rules. For instance, it may be tempting to accept the matches as probably correct but to follow up on the nonmatches because they may be erroneous due to defects in the source data and because the follow-up could yield better information. That is a correct procedure only if the matching rules are such that there is known to be a very high probability that the matches are indeed correct while many of the nonmatches may be erroneous. If, on the

other hand, the matching rules are such that the probability of error is about the same for matches and nonmatches, then both groups must be followed up if there is any follow-up at all.

It may be difficult to phrase the follow-up questions so that the maximum of new information is obtained. In most cases (except "possible matches") the interviewer should not be given the information already available and asked to verify it; that would be a temptation to just confirm it without checking, if checking is difficult (this is not a problem when the follow-up is done by mail). Nor should the follow-up usually be limited to asking again the same questions that were asked before; the answers would tend to be the same unless a different respondent happens to answer. Another follow-up problem, when current data are involved, is the need to get back to the respondent as soon as possible in order to minimize recall problems and the possibility that the study unit may move or cease to exist. That requires good planning and coordination so the data can flow from collection to matching to follow-up without delay.

Reliability of the results of an exact match project may be defined as the proportion of erroneous decisions, that is, false matches and erroneous nonmatches; or as the proportions of true matches detected and spurious matches included. In the special case of matching to eliminate duplication, reliability is expressed in terms of duplication left in the final file. The proportion of errors may be estimated in various ways. In some cases some independent information may make it possible to know or estimate in advance what proportion of the base records should be in the reference file (in a few cases this may be 100 percent, and a match rate of less than that would indicate either an inefficient matching procedure or an incomplete reference file-assuming that the records contain sufficient information for matching). Usually, if the files include some corroborating information, it will be possible to be practically certain about many matches; in some projects one may also be certain about many nonmatches. A sample of the remaining cases (and, for confirmation, a small sample of "certain" cases) can then be put through an additional round of searching with more thorough procedures, or more information can be obtained through field follow-up (by phone, mail, or interview). The information obtained in that

way for the sample cases can then be used to estimate error rates. Another possibility would be to obtain such estimates in advance through a pilot study.

.11 See References to B. and C.4; Neter el al., 1965.

As mentioned before (Section C.6), the estimation of error probabilities may be built into a matching procedure based on a mathematical model. Reliability could be improved by putting all (instead of a sample) of the records that are not either clearly matched or clearlv not matched through additional rounds of matching, or, if feasible, through a followup to get more information. But that would usually be very costly and would probably still leave a residue of cases for which it cannot be determined satisfactorily whether the base file records have no match in the reference file, or whether there is a matching record in that file which cannot be found because of defects of the available information. If the data are of

poor quality, the most complex routines and the most sophisticated computers will be of little use. Improvements in the reliability of matching applications can undoubtedly be made with greater certainty by concentrating on the quality of the input data, instead of devising complex and costly procedures to manipulate data of questionable information value.

F. Elimination of Duplication in One File

Although it is not included in the definition of an exact match used in this report, elimination of dupli-cation within a file is a special application of a procedure similar to exact matching. Instead of matching one file against another for possible matches the matching procedure must be set up to match each individual record with all other records in the file or all other records within blocks. If the file exceeds a few thousand records it will ordinarily be necessary to use blocking in order to control costs of computer matching or in order to control time and cost requirements of manual matching. Regardless of whether manual or

computer procedures are used it is usually best to block on two different factors and run the matching procedure twice. If manual matching is used to identify duplicate records for the same person, two different sort orders should be used. The first would be a completely alphabetic listing of the entire file and the second an alphabetic listing within zip code or city. The first listing will identify all of the complete duplicates (same name and address) and identify possible duplicates for which the name is exactly the same but address information has changed or may be in error. The second listing will enable matching of records with correct address information but name misspellings. A final step in the duplication removal might be to check common misspellings from the second listing back against the first listing. This procedure might enable the identification of possible duplicates which have common misspellings of the same name and addresses which are close together geographically.

CHAPTER III

Statistical Matching

A. Introduction

As noted earlier, the Subcommittee has defined a statistical match as a match in which the linkage of data for the same unit from the different files either is not sought or is sought but finding such linkages is not essential to the procedure. In a statistical match, the linkage of data for similar units rather than for the same unit is acceptable and expected. Statistical matching is a relatively new technique which has developed in connection with increased access to computers and the increased availability of computer microdata files. In a statistical match each observation in one microdata set (the "base" set) is assigned one or more observations from another microdata set (the "nonbase" set) ; the assignment is based upon similar characteristics. Usually the observations are persons or groups of persons, and the sets are samples which contain very few (or no) persons in common. Thus, except in rare cases, the

observations which are matched from the two sets do not contain data for the same person. This is in contrast to an exact match in which data are matched for the same person from two different sets. A statistical match can be viewed as an approximation of an exact match. (See Okner (1974) and Radner and Muller (1978) for papers which contain overviews of exact and statistical matching work.) Some statistical matching methods can be similar to exact matching methods. For example, the Census Bureau's Unimatch computer program (Bureau of the Census, 1974) has been used for both exact and statistical matching.12 Statistical matching methods can also be similar to techniques used to match data for other purposes, such as the "hot deck" allocation of data to non-respondents in household surveys (e.g., Spiers and Knott, 1970) or matched or paired sampling (e.g., Althauser and Rubin, 1969). Statistical matching as defined in this report differs from those other techniques because in a statistical match two different microdata sets are matched and (in almost all cases) the purpose is the addition of variables not present for any observations in the base set. In some cases those added variables can have the same definition as base set variables but contain less error. The study of statistical matching is still in its early stages. Many important theoretical and practical questions

about statistical matching have not been answered. These unanswered questions include:

1. How accurate are statistical matches?

2. For what purposes and under what conditions are the results of statistical matches sufficiently accurate?

3. What factors are important in determining the accuracy of the results of statistical matches?

4. What are optimal methods of statistical matching and how are those methods affected by the circumstances of the match?

5. Given a set of alternative statistical matching methods and a set of conditions, what is the relative accuracy of the different methods?

6. What are the best ways of handling practical problems such as those resulting from differences between samples and between the variables in the files?

7. How sensitive are the results of statistical matches to the assumptions made in carrying out the matches?

Of course, these questions cannot be answered here. We will merely try to summarize what has been done and what is known, and suggest directions for future work. In this chapter, a description of a simple framework within which statistical matching can be analyzed is followed by brief discussions of the steps carried out in making a match and two basic types of statistical matching methods. Then the history and development of statistical matching are sum-

.13 See Springs and Beebout (1976) for an example of a statistical match carried out using Unimatch.

marized, followed by brief discussions of general criticisms of statistical matching and errors in statistically matched results. Finally, a summary and conclusions are presented.13

B.A Suggested Framework for the Analysis of Statistical Matching Methods

In this section a brief summary of the theoretical steps involved in a typical statistical match will be followed by a somewhat more detailed discussion of those steps. An example involving household survey and income tax data will be used to clarify the concepts as the discussion proceeds. In summarizing the matching steps, we begin with a universe, "U," for which we want to make estimates of variables and their relationships to each other. We have two microdata sets, "A" and "B," samples which provide observations on the universe; each set contains some variables which are not included in the other set. We then define a hypothetical exact match result which we want the statistical match to approximate. However, we do not know the hypothetical exact match result; therefore we estimate it, either explicitly or implicitly, using whatever information is available. The appropriate matched pairs of units are then chosen in a way which minimizes deviations from the estimate of the exact match result.

1. Universe

We begin the detailed discussion of the framework by considering the universe U for which we want to estimate various relationships. U consists of a set of N units; for each unit there are values for R variables. By definition all information in U is error-free, and it is assumed that all information relevant to the estimates we want to make is contained in the R variables. U can be represented by an N x R matrix in which each of the N rows contains the values of the R variables for one unit.

2. Two Data Sets

We will assume that we have two microdata sets of observations on variables for units in U; these sets, A and B, are the sets we want to match statistically. A and B will be assumed to be samples from U. A contains n.A units, while B contains n.B units, where both n.A and n.B are less than N; n.B does not necessarily equal n.A. It will also be assumed that very few units from U appear in both A and B; A and B could be independent samples for which n.A/N and n.B/N are small. For example, set A might be the persons interviewed in a household sample survey for a given year, and set B might be a sample

of income tax returns for that same year. It will be assumed that A contains observations on k variables, while B contains observations on m variables. By assumption, both k and m are less than R, and all of the variables are contained in U. Some variables from U may be contained in both A and B, while at least some will be contained in only one set. The i.th unit in A, which will be denoted A.i, contains k observed variables, as shown below:

$$A.i = (a.il\ a.i2 \ldots a.ik)$$

Similarly, the i.th, unit in B contains m observed variables:

$$B.i = (b.il\ b.i2 \ldots\ b.im)$$

It will be assumed that at least some of the variables in A and B can contain errors, while in U they do not. Because of different error components, a variable from U which appears in both A and B can have different values in the two sets for the same underlying unit in U. For example, even if wage income were defined identically in the household survey and the tax return, the survey response might differ from the amount shown on the tax return.

3. Hypothetical Exact Match

At this point we have defined the universe and the two data sets which will be matched statistically. We will now define "C," a hypothetical data set which represents the result of an exact match (carried out without error) between A and B, if the underlying units represented in A were also represented in B. The set C is hypothetical because that exact match cannot be carried out. The exact match is impossible because very few of the units represented in A are also represented in B. By assumption C contains all k variables from A and all m variables from B, including their error terms. Because a statistical match is viewed as an approximation of an exact match, C is the data set which we try to approximate when we perform a statistical match..14 It is important to note that C is not necessarily unique. The form of C depends upon which data set, A or B, is taken as the base..15 We are assuming that A is the base set.

.13 Earlier versions of much of the material in this chapter

appeared in Radner (1974, 1977, 1979).

.14 There may be cases in which a statistical match is not an approximation of an exact match. For example, in some cases it might be useful to bias the match (relative to the exact match result) in order to adjust for underreporting of data and thereby avoid a postmatch adjustment step.

.15 One set can be used as the base set for part of the sample and the other set can be used as the base set for the rest of the sample. For

For the i.th, unit in A, the information in C will be denoted C.i, and can be expressed as follows:

C.i = (a.il a.i2 ... a.ik b*.il b*.12....b2.im)

= (A.i B.i*)

Using the previously mentioned example, Ci contains the survey response given by Ai and the data from the tax return filed by Ai. As noted above, that tax return does not appear in B, except in rare cases.

4. Estimate of Hypothetical Exact Match

When we actually want to make a match, we do not know C (i.e., we do not know B.i*). We therefore make (either explicitly or implicitly, depending upon the matching method) an estimate of C, called "L", using whatever information is available. This estimate is used in carrying out the match. Not all of the variables in B.i* need to be estimated. The estimated variables in B.i* (along with any constructed variables) will be used as "matching" variables; that is, they will be used to carry out the match. Estimated values can be obtained by assumption. For example, for a given A unit, it might be assumed that the value for a given B variable should be equal to the value for a given A variable (say, a.ll = bi*.ll). We could say that wage income in B should be identical to wage income in A. This would be valid if wage income were defined identically and had an

identical error pattern in A and B, which ordinarily is not true. When such an equality does hold, we have a special case in which, for those variables, the estimation of C is trivial. Estimated values can also be obtained by other means, for example, by regression techniques or by using information from an exact match between sets similar to A and B or from an exact match of subsamples of A and B. The estimates often vary in reliability for the different B variables. In some cases the estimates of B.i* are constructed in such a way that the distributions of the estimated variables approximate the distributions of the original B variables.

For the i.th unit in A, the information in L will be denoted L.i, and can be expressed as follows:

L.i = (a.il a.i2 ... a.ik b*.il b*.i2 ... b*.im) = (A.i B*.i)

Although we have shown all m vairiables estimate, as noted above, it is not necessary to estimate all of them. Using the continuing example, for each unit in A, L contains that unit's survey response data and estimates of some or all of the variables in the tax return filed by that A unit.

5. Statistical Match Result

We now introduce "M," the result of statistically matching sets A and B in some unspecified way. For the ill, unit in A, the information in M will be denoted M.i, and can be expressed as follows:

M.i = (a.il a.i2 ... a.ik bø.il bø.i2 ... bø.im = (A.i Bø.i)

In our example for each unit in A, M contains that unit's survey response data and the tax return data from the B unit assigned to that A unit in the statistical match.

It should be noted that in some cases, where sample weights differ, A units are assigned more than one B unit and sample weights are split so that the total weight of the A unit (and of the B units) remains unchanged.

It is not necessary for every B unit to be used in the match solution, and some B units can be used more than once in the solution..16 it follows from the definition of a statistical match that the m variables from each B unit are assigned as an entity.

In making a statistical match we choose among alternative solutions; each alternative solution is characterized by the

particular set of B units assigned and the particular A unit(s) to which each is assigned. We choose the solution in which M approximates L as closely as possible, in terms of the variables and relationships of greatest importance in the results of the match. This approximation can be viewed in terms of a "distance function." We can define in general terms a distance function, "D," which measures the distance (DM) of M from L. The distance function D is chosen according to the purpose of the match. Thus,

$$D.M = D(M, L/P)$$

where P denotes the purpose of the match..17 The statistical match solution which minimizes D.M is the optimal match result."

C. Applications of Statistical Matching

The vast majority of statistical matching work has been in the field of economics. The first statistical match in economics was performed at the Bureau of Economic Analysis of the U.S. Department of Com-merce in 1968 in connection with estimating the size

example, a tax return sample might be used as the base set for the high-income portion of a match (where it is the denser sample), while a household sample survey might be used as the base set for the rest of the sample (where it is the denser sample). In constrained matches (see p. 18), both sets are used as base sets for the entire sample.

.16 In some matching procedures every B unit is required to be used in the match solution, and used with its original sample weight. For exampl e, see Radner (1974) and Turner and Gilliam (1975).

.17 In this formulation, it is assumed that the distributions of the B variables in L approximate the distributions of those variables in C. If that is not true, then, in some cases, the formulation D.M = D(M,L,B/P) can be used since it might be desirable to approximate distributions from B.

.18 This is not meant to suggest that statistical matches should necessarily be carried out using distance functions; random selection within cells is one possible alternative.

distribution of family personal income. Another early match was performed at the Brookings Institution in connection with analysis of the tax system. More recent work has been done at Statistics Canada, Yale University (and the National Bureau of Economic Research), the Office of Tax Analysis of the U.S. Treasury Department, Brookings, the Office of Research and Statistics of the Social Security Administration, and Mathematica Policy Research. These matches were undertaken in order to construct more comprehensive and/or more accurate data bases from existing ones. Statistically matched files have been used to make estimates of the distributions of income, taxes, wealth, and the costs and effects of changes in government programs. Proposed uses include making estimates from "nested surveys" (Office of Federal Statistical Policy and Standards, 1978a) and the construction of microdata sets consistent with the sectors of the National Income and Product Accounts (United Nations Statistical Office, 1978). Most of the matches discussed here have been between household survey samples and tax return samples. Others were between

two household surveys, and between two files constructed from several types of data using exact matches.

1. Matching Steps

Several steps in actually making a statistical match should be mentioned here. First, if the populations represented by the two files differ, a "universe adjustment" might be needed. Second, a "units adjustment" might be needed if the units of observation in the two files differ (e.g., persons and tax units). Third, "matching variables", the variables in the two files which are used to choose the B set records to be matched with the A set records, need to be chosen. Ordinarily, matching variables are defined similarly in the two files and are highly correlated with important "nonmatching" variables. In some cases, matching variables are constructed as functions of one or more variables in the set. Fourth, whatever "linking information" exists needs to be identified. Linking information consists of information (or assumptions) about joint distributions of the matching variables in the two files in C. Fifth, that linking information is used in the construction of L (either

explicitly or implicitly). The construction of L includes the adjustment of values of matching variables (in one or both sets) to take account of differences in definitions and response and reporting error patterns,.19 as well as the construction of matching variables. Estimated values might be obtained by assumption. For example, as noted earlier, for a given A unit it might be assumed that the value for a given B variable should be equal to the value for a given A variable. We will call this assumption the "equality assumption." Estimated values can also be obtained by other means, for example, by regression techniques or by using cross-tabulations from an exact match between subsets of A and B or between sets similar to A and B. It is important to note that estimates of B set variables in L can vary in their reliability. Finally, in the "merging" step, the records from the nonbase set are chosen. Although many different methods have been used in this final step, several basic similarities can be identified. In most matches, both files have been separated into comparable subsets of units, or "cells." Within each cell, rules have been specified for the choice of one or more records from the nonbase file to be assigned to each record from the base file. The selection of the record often was based upon a distance function by

which a distance was computed between a given base set record and each potential match in the nonbase set. The distance was computed from differences between values of the matching variables in the two records. The potential match with the smallest distance ordinarily was chosen as the match.

2. Two Basic Types of Methods

Many different matching methods have been used. These methods will be separated into two principal types, "constrained" and "unconstrained," according to the extent to which the distributions of the nonbase set variables are used in the matching procedure. In a constrained match, every nonbase set record appears in the matched result and has a sample weight identical to its sample weight before matching..20 Thus, the distributions and joint distributions of nonbase set variables (as well as base set variables) are not changed by the match. In an unconstrained match, there is no such restriction on the nonbase set variableS..21 A constrained match can be viewed as choosing nonbase set records without replacement, while an

_____

.19 Such adjustments have been called "alignment" by Ruggles and Ruggles (1974).

.20 It should be noted that a nonbase set record can be matched with more than one base set record if the original sample weight of the nonbase set record is split among the base set records. It should also be noted that in practice the definition of a constrained match can be relaxed to include matches in which sample weights (in either file) are not identical before and after matching but can change only slightly (e.g., due to round-off error).

.21 Unconstrained matches could be separated into different types, for example, according to whether, and how, the distributions of the nonbase set variables are used in the construction of L.

unconstrained match can be viewed as choosing with Census, and the 1964 Tax Model (TM), an Internal replacement. A constrained match does not always Revenue Service sample of Federal individual income allow the best match for each base set record; thus, in a constrained match, on the average, the matches are not as close as can be obtained in an unconstrained match. However, in a constrained match, no reweighting error is added to the nonbase set information as ordinarily happens in an unconstrained match. A matched record will contain two sample weights-one from each file. In an unconstrained match, ordinarily the sample weight from the base set portion of the matched record is used in the results. Thus, the nonbase set information is reweighted. In a constrained match, the sample weights from the two files in a matched record will be the same.

3. History and Development of Matching Methods

Statistical matching in economics began as a solution to a specific problem faced by the Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce.22_improving the accuracy of and adding more detail to household sample survey income data (from the

Current Population Survey). The solution was a statistical match between the household sample survey and a sample of income tax returns. Such a statistical match was also the solution to a problem the Brookings Institution was interested in-putting a sample of tax returns on a family unit basis and adding nontaxable income types and nonfilers to the tax return data. However, BEA and Brookings chose quite different matching methods. The BEA and Brookings (MERGE-66) matches are the most important members of what might be called the first generation of statistical matches in economics. A second match carried out by BEA (the SFCC match described later) also belongs to the first generation. The other matches described here belong to the second generation. Those other matches took into account the results of and experience with the BEA and Brookings MERGE-66 matches.

a. Bureau of Economic Analysis, U.S. Department of Commerce, CPS-TM Match.23

The BEA CPS-TM match was between the March 1965 Income Supplement of the Current Population Survey (CPS), conducted by the Bureau of the tax returns. The purpose of the match was the improvement of the accuracy of CPS income amounts and the addition of tax return income detail to the CPS observations; the CPS was the

base set. There were some differences between the universes-some CPS persons did not file tax returns and some TM returns were filed by persons outside the CPS universe (e.g., persons abroad and some military personnel). The units in the two sets were differentpersons in the CPS and tax filing units in the TM. This was a constrained match; cells and ranking of records according to size of income amounts were used. The basic universe adjustment used was the esti-mation and elimination from the CPS of those who filed no tax return ("nonfilers"). After the definitions of the units in the two sets had been made roughly comparable by transforming CPS person units into tax filing units using small amounts of information from the 1963 Pilot Link Study (an exact match), the nonfilers were chosen as a residual. Units considered to have the lowest probability of filing were chosen to be nonfilers. There was very little empirical (exact match) linking information available. Matching variables were chosen on the basis of the (subjective) reliability of the assumptions regarding their joint distributions. After examination of the relevant overall (marginal) distributions (and taking into account the exact match information that did exist), it was assumed that the differential response error and differences in definition between matching variables in the two sets were important factors.

The ranking described below was used to take account of these factors. Cells were constructed for each matching variable. These cells were constructed in sequence, with the cells for the second variable defined within the cells for the first variable, and so forth. The variables used were (in order) marital status, wage and salary income, self-employment income, and property income. This formulation incorporated the linking information which suggested that the correlation between the CPS and TM amounts in an exact match carried out without error would be highest for wage and salary income, next highest for self-employment income, and lowest for property income, among the numerical matching variables. The specific assumption about the joint distributions of matching variables which was used was that units with approximately the same rank in the (conditional) distribu

---

.22 The Office of Business Economics (OBE) became the Bureau of Economic Analysis in 1972.

.23 Budd and Radner, 1969, 1975; Budd, 1971; Budd, Radner, and

Hinrichs, 1973; Radner, 1974.

tions of the specific variables in the two sets would be

for different years. The basic method was the sepamatched. That is, for numeric variables, the defini- ration of both files into cells and then, within cells, tions of cells were based upon rank rather than upon the absolute size of values. Although this assumption was consistent with the overall distributions in the two sets, it obviously was crude. The assumptions used also implied that, in each

cell, there would be the same weighted number of units in each set. In the final step in the match, observations in both sets were duplicated and their sample weights were split so that no sampling was needed and the overall distributions of all variables in both sets were preserved. One of the benefits of this technique was that it eliminated possible error arising from widely differing sample weights in the TM. A crude sensitivity analysis was carried out by comparing the constrained method results with the results of several versions of an unconstrained method (Radner, 1974). The BEA match gave a central role to differences between the matching variables in the two sets. Although this emphasis had its origin in the fact that the match had correction of income amounts as its purpose, differences between matching variables can be important factors in many matches, regardless of their purpose. BEA also emphasized the accuracy of the overall distributions of variables in the matched file. These two factors led BEA to use a constrained method.

b.Bureau of Economic Analysis, U.S. Department of Commerce, SFCC Match 24 A second early statistical match was also carried out in the BEA income size distribution work. This match was less detailed an d less important than the CPS-TM match described above, but it does

deserve mention as one of the earliest statistical matches. This match, performed in 1969, was between the statistically matched 1964 CPS-TM file (corrected for income tax return audit) and the Survey of Financial Characteristics of Consumers (SFCC). The SFCC contained income data for calendar 1962 and asset and liability data for the end of 1962 for roughly 2,500 households. The purpose of this match was the addition of data by which amounts of several income types not covered in the CPS-TM file could be assigned. Most of those income types were noncash types and most of the data added were asset data . This match was performed on a family unit (family or unrelated in dividual) basis, and was an unconstrained match. The unconstrained approach was chosen primarily because the two files contained data

Budd, Radner, and Hinrichs, 1973.

ranking the records in each file according to size of interest income. The specific SFCC record to be matched to a given CPS-TM record was the SFCC record with a corresponding ranking. Size of total money income, type of family unit, age, race, and major source of earnings were used as cell classifiers. These variables were chosen primarily because of their relationship with the asset types to be added to the CPS-TM file (interest income was used for the same

reason). SFCC records were reweighted so that, within each cell, the weighted numbers of records were equal in the two files. The records in both files were then ranked, within cells, according to size of interest income (from high to low); matching was carried out based upon that ranking. The matching did not involve the splitting of records as had been done in the CPS-TM match. Instead, for each CPS-TM record, the SFCC record which fell at a "selection point" In the series of cumulated sample weights was chosen. For a given CPS-TM record, the selection point was defined to be one third of the record's sample weight plus the cumulated sample weight of the CPS-TM record above it in the ranking. The highest ranking SFCC record whose cumulated sample weight was greater than or equal to that value was chosen as the match. For example, if the selection point was 6,000, then the highest ranking SFCC record with a cumulated weight of at least 6,000 would be the match.

c. Brookings Institution MERGE-66 25 MERGE-66 was between the Survey of Economic Opportunity (SEO) for income year 1966 and the 1966 Internal Revenue Service Tax File of individual federal income tax returns. This match was one step in the construction of a corrected and more detailed

microdata base for policy analysis, particularly tax policy analysis. The SEO was used as the base set; cells, ranges, and a distance function were used. This was an unconstrained match. Universe adjustments were made to both files: it was assumed that high-income (or loss) units were in the Tax File but not in the SEO, and some filers of tax returns were not in the SEO universe. The first step was the formation of cells in both sets based upon marital status, age, number of dependent exemptions, and income types received, including the major source of income; 74 cells were used. An acceptable range of major source income was defined for each SEO unit; this range was the

25 Okner, 1972.

SEO amount plus or minus two percent, with upper variables) to make those estimates. Sims defined X and lower absolute amount bounds. Then, for each variables, which appear in both sets, Y variables, SEO unit, each Tax File return which was both in the appropriate cell and with the acceptable major source range had a "consistency score" computed. This score, which was a simple distance function,-26 was based upon the correspondence of the existence of home ownership, property income, self-employment income, and capital gains in the two sets (some of that information was estimated in each file). The group was then narrowed down by including only the 25 percent of the group with the highest consistency scores. In addition, a minimum absolute consistency score was required. If this top 25 percent group was "large enough," then a Tax File return was selected randomly, with the probability of selection for each return proportional to its weight. If the eligible subset was "too small," then the major source income band was widened and the whole process was repeated. The basic procedure was essentially to treat the SEO units one at a time and to define a small

subset of the Tax File from which one return would be drawn randomly. Thus, the one best match for each SEO unit was not I identified; the final selection was random. The equality assumption was used for all variables, both reported and constructed. The basic approach used in the construction of L (the estimated hypothetical exact match) was what might be called a " modal" one; the most common value of the variable was used in L. MERGE-66 can be compared to the Census Bureau's hot deck allocation procedure. The hot deck procedure, which can be thought of as the state of the art" of record matching in economics (ot . her than exact matching) prior to the advent of statist I cal matching, resembled an unconstrained match with no differences between matching variables. M ERGE-66 was similar to the hot deck method in that respect. In contrast, the BEA match was a marked departure from the hot deck precedent.

d. Christopher Sims' CommentS27 A word should be said about Christopher Sims' two early "Comments" on MERGE-66 and other matching procedures. Sims formulated the statistical matching problem as the estimation of the joint distributions of variables which appear in only one of the sets being matched (non-common variables), using variables which appear in both sets (common

In this distance function, the higher the value the better the match. This is the opposite of distance functions described earlier in which lower values were better. Both types are referred to as distance fuinctions in this report.

27 SIMS, 1972, 1974.

which appear in only one set, and Z variables, which appear only in the other set. The X variables in the two sets are then matched, and estimates of the joint distributions of Y and Z are obtained. Sims interprets the MERGE 66 and other procedures to assume that Y and Z are independent conditional upon X. This formulation suggests conclusions regarding the accuracy of statistically matched sets. Sims' formulation of the statistical matching problem has been quite influential. However, it should be noted that that formulation applies to a special case of the generalized statistical matching problem. Two limitations on the applicability of his formulation should be mentioned. First, Sims gave little attention to the joint distributions of the matching variables in the two sets. In his formulation, in effect he assumed that the equality assumption was valid (although he did mention the adjustment of matching data). However, the separation of variables into X (variables which appear

in both files), Y (variables which appear only in one file), and Z (variables which appear only in the other file) is frequently not applicable. In many cases the variables used to match on (X's) are not strictly comparable; that is, they differ in definition or error component (e.g., response error), or both. In general, there can be a range of degree of comparability between pairs of variables in the two files. Pairs of variables are chosen as matching variables when, as a necessary condition, information about the joint distributions of those variables (in an exact match carried out without error) is known or can reasonably be inferred. When the matching variables are chosen, the variables are separated into matching and nonmatching variables, but the matching variables often differ in the reliability of the information available about their joint distributions. These differences can be reflected in the matching method. The second limitation is that the purpose of the match is not always only the estimation of the joint distribution of non-matching variables in the two files. In many matches the matching variables from the nonbase set have been used in the results of the match. Where tax return files have been used, the matching variables from the tax return data have usually been used in the results of the match. This has been

done primarily because it was desirable to use the entire set of tax return variables as an entity. However, it should be noted that where the matching variables in the two files differ in definition or in the amount of error they contain, it can be useful to use

the matching variables from the nonbase set in the results even if the use of the nonbase set data as an entity is not crucial. For example, some nonbase set matching variables might contain less response error.

e. Statistics Canada SCF-FEX Match28 The Statistics Canada match was carried out between two Canadian microdata sets, the Survey of Consumer Finances (SCF) and the Family E penx diture Survey (FEX), which contain data for 197 . 70. The purpose was the

addition of expenditure data to the SCF. This match had the advantage that both microdata sets were obtained using the same sampling frame, the Canadian Labour Force Survey. Thus, both the universes and the definitions of units were identical. In addition, many of the variables in the two sets purposely were defined identically. The approach was influenced primarily by MERGE-66. This was an unconstrained match, using the SCF as the base set. Cells and a distance function were used, as was the equality assumption.

The first step in this match was to use multiple linear regression analysis to determine, given the purpose of the match, which variables should be used as matching variables, and how much weight should be given to each of those variables. This step represented an attempt to make the choice of matching variables and their relative importance more objective. This attempt was in contrast to both the BEA and MERGE-66 matches in which those choices were almost entirely subjective. In the regressions, the independent variables (income and demographic characteristics) were variables which appeared in both sets. The dependent variables chosen appeared only in one set and were important to the results of the match; the SCF dependent variables were asset and debt information, and the FEX dependent

variables were expenditure information. Both sets were separated into four subsets based upon home ownership and type of consumer unit prior to the running of the regressions. Once the matching variables had been chosen, they were separated into "mandatory" and "desirable" variables. The mandatory variables (which were categorical variables) were used to partition the sets into cells. Following the precedent of the MERGE-66 consistency scores, "union scores" were computed for desirable variables; this was a distance function. Different maximum point totals were assigned to different linking variables on the basis of the regression results; the greater the variable's explanatory power, the greater its maximum point total. For

'Alter, 1974.

example, "no discrepancy in amounts of major source income" was worth 40 points, while "no discrepancy in total income" was worth 30 points. The Statistics Canada technique differed from the MERGE-66 technique by assigning different point values to discrepancies of different sizes; the MERGE-66 version was "all or nothing" in concept. A ranking procedure was used in the merging step. Records

in both sets were ordered according to size of income within the mandatory cells. Then the first FEX record with at least a 95 percent union score was matched with the relevant SCF record. Some SCF records were not matched in the first run and the subsequent runs which were necessary because of the effect of file sequence. Further runs were made with the minimum acceptable consistency score lowered. Finally, several variables were changed from mandatory to desirable so that all SCF records could De matched. The FEX records were used with replacement. The ranking procedure produced biases, which are commented on in Alter (1974). Statistics Canada also presented data regarding the quality of the matching. For example, the correspondence of codes of variables which were used as desirable matches was checked. In summary, the Statistics Canada match contained three responses to the earlier matches: (I) an attempt to make the choice of matching variables and their relative weights more objective; (2) a refinement in the use of distance functions by relating the distance (or union score) to the size of the deviation (discrepancy) and (3) an emphasis on attempts to assess the quality of the matching. f. Yale University (and National Bureau of Economic Research) 29 The Yale group was interested in devising

a generalized statistical matching procedure which can be applied efficiently to very large microdata sets (i.e., those containing several million observations). In this respect, the Yale work differed from that carried out at BEA, Brookings, and Statistics Canada. In those matches the procedures were tailored to the particular sets being matched, sets which were not very large. The Yale approach can be viewed as having its origin in the comments by Sims. An important part of the Yale work is an attempt to make the selection of cells more objective. The procedure contains two important parts, the "sort-merge strategy" and the estimation of "I(X)" regions. The sort-merge strategy is a technique for implementing the use of cells which is particularly appro-29 Ruggles and Ruggles, 1974; Ruggles, Ruggles, and Wolff, 1977;

Wolff, 1977.

priate for microdata sets with large numbers of distributions of the non-common variables are disobservations. In each file, for each of a set of match- similar. Thus, when the chi-square test shows a ing (or "common" or "X") variables, each observation is assigned a set of sort tags. These sort tags represent cells in the variable; more detailed (narrower) cells are nested within the broader cells. If there are n levels of detail for the cells, and m matching variables, then each observation will have nm sort tags (cell codes) assigned to it. The purpose of having different levels of detail is to ensure a match for every A file observation. An A file record is matched with a B file record with identical sort tags for all matching variables at the most detailed cell level possible. The procedure allows B set records to be used more than once, or not at all; thus, the procedure is of the unconstrained type. Because both files only need to be sorted once on the basis of these nested sort tags (with the least detailed set as the primary sort), the costs of

matching large data sets are held down. In most cases, the estimates of the I(X) regions define the cells which correspond to the sort tags. The estimation of the regions follows the lines suggested in Sims (1972). The I(X) regions are ranges of the matching (X) variables for which the distributions of the non-matching variables are significantly different. Matching takes place within corresponding I(X) regions in the two sets. In this technique the X (matching) variables are used only as intermediaries in the estimation of the joint distributions of the non-matching variables in the two sets. It is in this view of the matching problem that the Yale procedure follows from Sims. The estimation of the I(X) regions is an attempt to find an objective way to construct cells for matching, a goal which was similar to Statistics Canada's. Chi-square tests and the size of correlation coefficients between two distributions are used to estimate the I(X) regions. To make these estimates, observations in adjacent ranges of any common variables are treated as though they belonged to different samples. A chi-square test is then applied to test whether the distributions of the non-common variables in the two ranges of the common

variable are significantly different. If they are not significantly different, the two ranges can be combined. If they are significantly different, each of the ranges is split into two parts and those parts are tested in a similar manner. Because of the sensitivity of the chi-square tests to the number of observations involved, those tests are modified by examining the size of the correlation coefficient between the distributions which are being tested. If the correlation coefficient is low, then the

significant difference and the correlation coefficient is low, the ranges are not combined. By varying the significance levels for these tests, the different levels of detail and hence different numbers of cells are defined. It is in this way that more detailed sets of cells are nested within less detailed cells.

Wolff (1977) describes an application of the Yale method, the construction of the "MESP" database, which is the result of three statistical matches and two sets of imputations. That file, which contains asset and liability and demographic information for a sample of roughly 60,000 households, was con-

structed to serve several purposes; Wolff used it to estimate household wealth distributions. No single database contained the data necessary to make those estimates. The first statistical match in the construction of this file was between the 1969 IRS Tax Model and an augmented version of the 1970 IRS Tax Model of individual returns. Although the 1969 Tax Model was the file of most interest, the 1970 file contained race and age data (matched in from SSA records in an exact match) and more detailed data on itemized deductions which were not in the 1969 file. The 1969 file was the base file in this match; data were transferred from the 1970 file to the 1969 file. Broad cells based upon return type, sex, age exemptions, and number of children were used; the Yale method was applied within those cells. Size of adjusted gross income (AGI) and the major components of AGI as percentages of AGI, and total deductions were used as matching variables. Differences between AGI in the files arising from the fact that the data were for different years were handled by using percentile ranks. The second match, which was the basic match, was between the result of the first match and the 1970 Decennial Census 15 percent Public Use Sample (PUS). The PUS file was the base file, and detailed information on income from

assets along with other information was transferred to the PUS file. Broad cells based upon return type, sex, race, and age were used. The matching variables used within those cells were total income, wage and salary income, self-employment income, number of children, and home ownership status. Total income and business and professional income were matched according to percentile rank in order to adjust for lack of comparability. The third match was between the 1970 15 percent PUS and the 1970 5 percent PUS; the 15 percent

file was the base file. The 5 percent file contained data on stocks of some consumer durables which were not in the 15 percent file; those data were added to the 1 5 percent file. Marital status, age, sex, race, and home ownership status were used as broad cell variables. Matching variables within those cells were total family

income, wage and salary income of the family head, property value, wage and salary income of the spouse, number of children, and home ownership status. Using the third match, Ruggles, Ruggles, and Wolff (1977) reported on tests of the accuracy of the matched results. Several regressions were run using both original and imputed variables, and Chow tests were performed on the regression coefficients. In 40 of the 42 Chow tests performed there were no significant differences between coefficients estimated using original sample variables and those estimated using original and imputed variables. Ruggles, Ruggles, and Wolff concluded that the statistically matched results were reliable enough for many applications.

g. Office of Tax Analysis, U.S. Department of the Treasury 30

The statistical matching work being carried out at the Office of Tax Analysis (OTA) is a logical extension of the constrained method first used by BEA. OTA's emphasis in the methodology is on the development of a technique to implement constrained matching. OTA uses a linear programming approach; the solution to the matching problem is to treat it as a transportation model. In theory, a distance function is minimized simultaneously for all units, given the constraint that

each input record in each file must appear in the matched file with its original sample weight. In practice, efforts have been made to reduce the number of computations needed. For example, subsamples of the input files have been used, and files have been partitioned into subsets prior to the minimization. Differing sample weights between and within samples are handled as an integral part of the procedure. In the merging step, units in each set have their sample weights split and many are matched with more than one unit in the other set. This splitting is similar to that used in the BEA CPS-TM match, except that in the OTA case simultaneous minimizations of distances rather than ranking is used to determine the splits. The equality assumption has been used. OTA has applied its method to subsamples from

30Turner and Gilliam, 1975; Barr and Turner, 1978a, 1978b, 1979; Wyscarver, 1978.

the 1973 Statistics of Income and CPS files and subsamples from the 1975 Statistics of Income and 1976 Survey of Income and Education files. In the latter match, age, race, sex, tax schedule, number of exemptions, adjusted gross income, wages and salaries, business

income, and property income were used as matching variables; some information about the correspondence of the values of matching variables in the matched file has been presented (Barr and Turner, 1979). (Detailed descriptions of these matches are not available at this time.) Kadane (1975, 1978) has done theoretical work in connection with the OTA method. Sims (1978) has commented on Kadane's work.

h. Brookings Institution MERGE-7031 The MERGE-70 file was constructed for analysis of the tax and income distributions. The match was carried out between the March 1971 CPS and the Internal Revenue Service's 1970 Individual Income Tax Model. The method was an unconstrained type, and consisted of the use of a distance function within a range and cells. Universe adjustments were made so that parts of both files were not matched. In general, the CPS was used as the base file. The basic procedure consisted of making the files as "comparable" as possible, then constructing pseudo tax data for CPS units, and choosing a tax return from the Tax File for each CPS unit. A substantial amount of adjusting for universe and unit

differences was made. Tax units were constructed from CPS data, and CPS units which were estimated not to have filed were omitted from the portion of the file to be statistically matched. Three marital status groups were allowed: joint, head of household, and single. The Tax File had had age, race, and sex of filer added from SSA earnings records in an exact match (except for high-'income records) in order to increase the number of matching variables. Both files were partitioned into records which would be matched statistically and those which would not. For example, units in either file with large total income or a large loss in any income component, or both, were not matched. Persons living abroad and some armed forces members were eliminated from the Tax File. Separate and surviving spouse returns were also dropped from the Tax File; this was done because no CPS tax units having separate or surviving spouse returns were constructed. Some adjustments to income amounts were made prior to matching. In the CPS, amounts for specific income types were estimated from amounts

m Armington and Odle, 1975.

for broad income types and some property income was added. Audit correction factors were applied to Tax File income amounts prior to matching. The basic cell classifier used was whether wage and salary income was the primary income source. This classification was used to separate the file into wage and non-wage subfiles; different matching rules were used for those two subfiles. For the wage subfile, both files were partitioned into six groups based upon size of wage and salary income. Within each group, for each CPS unit the Tax File return closest in amount of wage income and the 37 returns above and below in the ranking by size of wage income (and within 20 percent of the CPS amount) were eligible for matching. Non-wage income fields were required to differ by less than $1,501

and CPS joint returns could only be matched with joint returns. The distance was then computed for each eligible pair and the Tax File record with the smallest distance was chosen as the match. The distance function included number of dependents, exemptions, sex, age and several income types. Each variable was assigned a weight in the distance function. The non-wage subfile in each set was partitioned into three groups, based upon the size of the total income variable. Several restrictions designed to avoid assigning too much of income types not in the CPS were used, The distance function for this sub-file used dependents, exemptions, sex, age, amount of total income, and presence of wages, dividends and interest, business, farm, rent and royalty, and miscellaneous income, with each assigned a weight. The ranking used to determine the eligible records was based upon total income. Apparently, the basic procedure did not use the sample weights from either the CPS or the Tax File. The distance function was of the following form for the ill, pair of variables:

$$D_i = \sum_{I} a_i - b_i$$

I ai + I bi

where ai is the A set value and bi is the B set value for the particular B record being considered. The distance for the B record was the weighted sum of the distances for variable pairs. After the initial match, the matched tax return data, using CPS sample weights, were compared with Tax File data. Problems were identified in two areas in the wage subfile. First, it was found that there were too many returns with large negative AGI. This problem was solved by rematching nine records. It was also found that there were too many returns with high capital gains. Apparently this

problem resulted from the fact that the sample weights in the highly stratified Tax File were not taken account of (this perhaps explains the negative AGI problem mentioned above.) This problem was solved by rematching units with large capital gains using a stratified subsample of returns with large capital gains for the matching. Data in the complete matched file (using CPS sample weights) were compared with the corresponding Tax File data and significant differences were found only for capital gains. The aggregate amount of business

income in the final file was also a problem. The distribution of distances for matched records was also examined. i. Office of Research and Statistics, Social Security Administrati on 32 The two input files to this statistical match were the 1973 Exact Match file (EM) and the Augmentation File (AF). The EM was constructed

by per-

forming an

exact match

between the

March 1973

Current

Population

Survey, SSA

earnings and

demographic

data for

1972 and a

limited

amount of

Internal

Revenue

Service

information

from federal

individual

income tax returns for 1972. The AF was constructed by performing an exact match between a subsample of the Statistics of Income federal individual income tax return file and SSA earnings and

demographic data. The AF contained detailed income tax return data, including tax liabilities, which were not present in the EM. The purpose of the match was the addition of income tax liabilities and more income

detail to the EM. The resulting file will be used for income and tax distribution analyses and for policy simulations. The EM contained roughly 42,000 records with tax return data, and the AF contained about 95,000

records. In this statistical match, for each EM record which contained income tax return data, the AF was searched for the observation which was thought to most closely resemble the tax return actually filed by

that EM unit (and that unit's SSA data). An unconstrained method was used. The match was made using cell categories and ranges, and a distance function to choose the best match within a cell and

range

combination.

The AF

records were

used with

replacement.

Twenty-two

variables

were used to

make the

match.

These

variables

either were

important

themselves

in the

results of

the match or

were

associated with important variables which could not be matched on. The following 10 variables were used to classify both files into cells: number of taxpayers; sex; race;

Radner, 1977, 1978; also see Appendix 11.

marital status; number of dependent exemptions; ments. About 83 percent of the records had two of type and size of earnings (SSA) ; existence of wage the three segments added while roughly 15 percent and salary, interest, and dividend incomes. Age and adjusted gross income were used as ranges around the EM value. Nineteen variables, including most of those used as cell classifiers, were used in the distance function. These nineteen variables included the existence of several income types, such as self-employment and capital gains. In general, the AF record with the lowest computed distance was the match chosen. If no acceptable match was found using the most detailed cells, the cell categories were made less restrictive and distances were computed; this process was repeated through four "levels". Most of the variables used to make the match were defined (almost) identically and would be expected to have (almost) the same reporting error pattern in the two files. Thus, the equality assumption was used. The distance function consisted of the sum of weighted distances between the AF values and the corresponding EM

values, for the nineteen variables. The importance and comparability of the matching variables were reflected in the weights applied to the distances. The distances were functions of the differences between AF and EM values.

j. Statistics Canada COC and MCF Matches:':' Statistics Canada has recently carried out two statistical matches combined with sample surveys as an alternative to censuses. The censuses were not undertaken because of cost considerations and the desire to keep respondent burden as small as possible. In these matches, tax return data were used to supplement survey data on businesses. The Census of Construction (COC) and the Motor Carrier Freight (MCF) survey were the surveys which were matched with the tax return data. These were unconstrained matches. This summary will focus on the COC match. In the COC match, a sample of roughly 41,000

businesses was constructed which consisted of the following types of records: Percent of Observations Basic tax return data only 83 Basic and secondary tax return data only 5 Basic tax return and survey data only 10 Basic and secondary tax return data and survey data 2

The objective was to assign the missing segments of data so that all records would have all three seg- Colledge et al., 1979.

had one segment of data added. This work differed from the majority of the other matches described in this chapter in two basic ways. First, some of the observations began with the different segments of data exactly matched. In fact, all of the data available for the secondary tax return and survey segments were exactly matched with the data from the basic tax return segment. Roughly two percent of the records did not have any segments assigned because all three segments were exactly matched. Second, three (rather than two) different sets of data were involved in the matching. In effect, this work contained two basic statistical matches. One was between records with secondary tax data present and those with those data absent; the other was between records with survey data absent and those with those data present. In the latter match the survey

segment was assigned in several parts, rather than as a unit. Each of the basic matches was similar to a "hot deck" nonresponse allocation in that one file (donor) contained all of the relevant segments of data, while the other file (candidate) had one relevant segment missing. Province (or region), standard industrial classification, and presence of wage and salary income were used as cell variables. Records in both files were ranked by size of gross business income within each cell. For a given candidate record, the nearest five donor records above and below it in size of gross business income were eligible for matching. A distance function was then computed for those ten donors and the donor with the smallest distance was chosen. In general, the absolute value of the difference between the logarithm of total expenses in the two records was used as the distance. Statistics Canada attempted to assess the sensitivity of the results by carrying out a small simulation. Sampling bias and sampling rate were examined in that simulation.

k. Mathematica Policy Research Mathematica Policy Research has carried out several statistical matches in connection with policy analysis performed for various agencies of the Federal government. Completed work includes matches between: a subsample of the 1970 Decennial Census Public Use Sample and the 1973 Aid to Families with

Dependent Children Survey (Springs and Beebout, 1976) ; the March 1975 Current Population Survey and the Survey of Household Characteristics, a survey of food stamp administrative records, (Beebout, Doyle, and Kendall, 1976); the Michigan Panel

on Income Dynamics (MPID) and the Nationwide Personal Transportation Survey (NPTS) (King, 1977); and the statistically matched MPID-NPTS file and a subsample of the 1970 Decennial Census Public Use Sample (King, 1977). These matches were carried out using cells and a distance function; a modified version of the Unimatch program was used. In most of these matches, a combination of subjective choices and regression analysis was used in specifying the matching variables and the relative importance of those variables. Other statistical matches planned by Mathematica Policy Research

include those between: the Survey of Income and Education and the Health Interview Survey (Pappas, 1979), the 1974 Survey of Purchases and Ownership (SOPO) and the 1976 Annual Housing Survey (AHS) (Hollenbeck, 1978), and the statistically matched SOPO-AHS file and the Survey of Income and Education (Hollenbeck, 1978).

1. Other Statistical Matches A statistical match carried out by Richard Rockwell between the 1970 Decennial Census Public Use Sample and a Survey of Economic Opportunity file is mentioned in Ruggles and Ruggles (1974). Five variables were used to define 288 cells and matches were made within those cells using three additional variables. Raymond Pepe performed a statistical match at the Bureau of Economic Analysis between the BEA 1964 Income Size Distribution File and the 1960-61 Consumer Expenditure Survey.34

D. Criticisms of Statistical Matching

There have been several published exchanges which have focused on criticisms of particular matching methods. For example, see Okner (1972), Sims (1972), Peck (1972), and Budd (1972); Ruggles and Ruggles (1974), Alter (1974), and Sims (1974); Kadane (1978) and Sims

(1978); and Barr and Turner (1978a) and Goldman (1978). Aside from the criticisms of specific matching procedures and matches contained in those exchanges, there have been several published criticisms of statistical matching in general. Sims (1972) objected to the construction of artificial samples by statistical matching. He argued that the artificial sample would have the correct joint distribution only if the sets of matching and nonmatching variables were mutually independent, and that that independence would be present rarely, if ever. Sims stated that if the

34 This match was carried out in connection with a Ph.D. dissertation to be filed with the Pennsylvania State University Graduate School. nonmatching variables were independent conditional on the matching variables and the regression functions between matching and nonmatching variables only changed slowly in the relevant ranges (i.e., between the values of matching variables matched in the two files), then the statistically matched sample would approximate the distribution of a true sample. He felt that those conditions are rarely, if ever, fulfilled. Sims

(1978) stated that the objectives of the statistical matches which have been carried out could be fulfilled better by other means. Specifically, he suggested computing histograms from the two original data sets. Fellegi (1978) expressed caution about the use of statistical matching because the accuracy of the joint distributions produced in the matched file is not known. According to Fellegi, statistical matching is based upon untested assumptions; he called for testing of statistical match results.

E. Types of Errors in Statistically Matched Data

Very little work on the errors present in the results of statistical matching has been done. (See Sims (1972), Wolff (1974), and Ruggles, Ruggles, and Wolff (1977) for examples of work that has been done.) Given this lack, we will merely attempt to identify several types of errors which can arise in statistical matching, assuming that the matching is done in an optimal way. "Error" is defined as the difference between data from an exact match of the two files (carried out without mismatches or nonmatches) if such a match were possible, and the data from the statistically matched file. In Chapter II, Type

I and Type 11 errors were discussed in connection with exact matching. Those categories of error are not applicable to statistical matching since the linkage of records for the same unit in both files rarely occurs in a statistical match. Thus, all or almost all linkages in a statistical match are mismatches in the terminology used for exact matches. However, both statistical and exact matching share the concept of error in the results of the matching (as contrasted with error in the matching itself). The error in the results of both statistical and exact matching can be viewed using the results obtained from a hypothetical exact match carried out without mismatches or nonmatches as the standard. In statistical matching a distinction should be made between "gross" error and "net" error. Gross

error refers to error on an individual record basis differences between values of specific variables). Net error refers to the error in some result in the matched file (e.g., the joint distribution of a pair of nonmatching variables in the two files). Offsetting errors can be an important factor in net error; gross error in different records can be offsetting. In some cases gross error could be substantial while net error was unimportant. However, if net error is substantial, then gross error must also be substantial. The error discussed below is gross error. Although net error is the more useful concept, it is very difficult to make statements about net error given the lack of research in this area.

The following three sources of gross error can be identified. First, because of lack of comparability between matching variables in the two sets (i.e., the variables are not defined identically and/or have different error patterns), we cannot know with certainty the values of the matching variables that we are searching for in the nonbase set. Second, even if we knew those values with certainty, often we could not find a nonbase set record with such values because the nonbase set is a sample which ordinarily does not contain the true match. Third, even if we could find a nonbase set record with

such values (assuming it is not the true match), the values for nonmatching variables in the nonbase set probably would differ from the true values because those nonmatching variables are not "completely explained" by the matching variables. It should be noted that these three sources of error can be offsetting. For example, we could be searching for a value which was too high, and find one which was lower than the value searched for.

A simple example might clarify the concepts. Assume that the match is between two sample surveys of white males and that no person was interviewed in both surveys. Assume that in survey A, persons were asked age, wage income, and years of education; and assume that in survey B, persons were asked age, wage income, and total income. The aim of this match is the estimation of the joint distribution of years of education and total income; age and wage income are used as matching (intermediary) variables.

Initially it will be assumed that total income is "completely explained" by age and wage income; that is, if a B unit which has the correct age and wage income is chosen, then the value for total income will be correct. It will also be assumed that age and wage income are defined identically and have

the same error components in the two surveys. Using the example of a 43 year old with $12,541 wage income and 12 years of education, sources of error will be examined. Under the above assumptions, it is known with certainty that we are looking in the B set for a 43 year old with $12,541 wage income. Because B is a sample, it is quite likely that no such record exists in B-thus, the fact that B is a sample which does not contain the true match is one source or error. However, if a B unit which is close to those values (e.g., 45 year old with $12,503 wage income) can be found, then the estimate of total income might be close to the true value. But, let us now assume that age and wage income are not defined identically in the two surveys and that the response error patterns in the two surveys can differ. Under these assumptions, we cannot say with certainty what values for age and wage income we are looking for n B-this lack of comparability between matching variables is another source of error. One assumption which has been used is that the values in B are identical to the values in A. In that case, even if we found a B unit with those values, it is likely that the value for total income would be incorrect, and it might not even be close to the true value.

NOW let us assume that total income is not completely explained by age and wage income-this is another source of error. Under this assumption, even if we know with certainty the values of age and wage income we are searching for, and even if we find a B unit with those values, the value for total income might be far from the true value. In matches made in the real world, we ordinarily have all of these sources of error; in different matches the relative importance of the difference sources can vary. One other specific source of error should be mentioned because it is frequently present-differences between the populations represented by the two sets. For example, if the B set contains units which are not represented in the A population, and the joint distribution between matching variables and total income differs between those units and the A population, then B set units not represented in the A population, if chosen in the match, can produce estimates of total income which are far from the true values.

F. Summary and Conclusions

Many different statistical matching methods have been used. No consensus regarding the best method

or methods has developed; both constrained and

CHAPTER IV

Findings and Recommendations

A. Findings

1. Definitions of Exact and Statistical Matching

Although the terms "exact" and "statistical" matching have been used frequently in the literature, the Subcommittee knows of no generally agreed upon definitions of these terms. For purposes of this report, the Subcommittee has defined a match as a linkage of records from two or more files containing units from the same population. It has defined an exact match as a match in which the linkage of data for the same unit (e.g., person) from the different files is sought; in

exact matching, linkages for units that are not the same occur only as a result of error. The Subcommittee has defined a statistical match as a match in which the linkage of data for the same unit from the different files either is not sought or is sought but finding such linkages is not essential to the procedure. In a statistical match, the linkage of data for similar units rather than for the same unit is acceptable and expected. Statistical matching ordinarily has been used where the files being matched were samples with few or no units in common; thus, linkage for the same unit was not possible for most units. The definition of a match used here excludes such record linkage techniques as the "hot deck" allocation of values to nonrespondents in surveys because those techniques are considered to involve only one file.

2. Usefulness of Matching

Matching of microdata sets is very useful for research and statistical purposes. Through the use of matching, it often is possible to carry out analyses or make estimates at a lower cost or in a shorter time than by alternative methods (e.g., a sample survey). In some cases, matching is the only feasible way of doing the

research. Analyses or estimates obtained through matching sometimes are more reliable than those obtained in other ways (e.g., for some kinds of information, matched administrative record data are more accurate than survey responses). Also, matching often leads to a reduction in response burden. The specific uses to which matching for research and statistical purposes has been put include the following: the addition of more variables to make possible analyses which otherwise could not be done or to enrich analyses with more variables; the evaluation of data, in which initial variables are compared with added variables or with additional reports on the same variables; evaluation of coverage; construction of more comprehensive lists.

3. Applications of Exact and Statistical Matching

Exact matching has been used for all of the purposes listed in 2. above. For many purposes statistical matching is inherently unsuitable. For example, analyses of census or survey coverage using record checks require matching of the same units (e.g., persons). Also, the construction of cumulative health histories and tests of

treatment effects ordinarily require exact matching. If we want to comare the earnings of persons who have had a given

. with those who have not, an exact

training program match between a list of trainees and earnings records is needed. A statistical match between those two data sets would not give useful results unless the earnings observations could be separated into persons who had been trained and persons who had not. However, statistical matching has been used for several purposes. One is the construction of microdata bases for policy analysis (e.g., for the analysis of the effects of current laws and programs and the estimation of the costs and effects of proposed changes). Another purpose is the construction of estimates of the distributions of various economic variables (e.g., income, taxes, and wealth). Other purposes involve the addition of variables to make

possible or broaden the analyses to be performed. Statistical matching has rarely, if ever, been used to combine microdata files which could be combined using an exact match.

4. Comparison of Errors

When there is a choice between statistical and exact matching, estimates of parameters of the joint distributions of variables in the different files will almost certainly have less error if based upon exact matching, To the extent that records forthe same person are successfully linked in an exact match, such estimates will be based upon data sets in which all the values of the variables for each person are in fact for that person; whereas in statistical matching, most or all of them are for a person with similar characteristics but not the same person. Error in exactly matched data has been studied and its effect can be estimated in many cases. On the other hand, little is known about the nature and extent of the errors present in data resulting from a statistical match. Most of

the literature on statistical matching has consisted of descriptions of matches performed, with little evidence presented on the errors in the matched results. These errors are very difficult to estimate. Thus, given what is known at this time, statistical matching is not a satisfactory substitute for exact matching in most cases.

5. Comparison of Relative Risk of Disclosure and Potential for Harm to Individuals

Confidentiality problems clearly are greater for exact matches than for statistical matches, for two reasons. First, if personal identifiers are used (as they usually are in exact matching), units (e.g., persons) must be identified, at least at some stage of the matching. Second, in an exact match (assuming that the true match is found) the matched file contains more information regarding the person than either of the original files. Thus, there is an increased probability of a record in the matched file being identifiable even after the removal of the personal identifiers. However, in most applications that probability is still very small. These problems exist to a lesser degree in the case of statistical matching. Protective measures against disclosure can be taken in

both cases, but for exact matches they may entail greater expense and/or some loss of information. The potential for harm to individuals resulting from inadvertent disclosure of identifiable records depends on the amount and sensitivity of information in those records. Since exact matching increases the amount of information in individual records, it can increase the potential for harm resulting from inadvertent disclosure. However, the Subcommittee believes that the Federal agency exact matching projects for statistical purposes which it has reviewed (see Appendices I and 111) have been carried out with sufficient safeguards to insure a very small risk of harm to specific individuals resulting from inadvertent disclosure of information about them in the matched files. No case has come to the Subcommittee's attention in which individuals have been harmed or have alleged harm resulting from such disclosures of individually identifiable records. The Subcommittee cannot, of course, assert that this has never happened or that Individuals have never been harmed as the result of the publication of statistical information about the population subgroups to which they belong. If the potential for harm from

publication

I publication of subgroup data were to be completely eliminated, the publication of Federal statistical data, whether or not based on matched records, would be severely curtailed.

6. Legal Obstacles to Exact Matching[

hrttab}Over the past 5 years, there have been significant changes in the laws and regulations pertinent to exact matching of records for statistical and research purposes. New laws, especially the Privacy Act of 1974 and the Tax Reform Act of 1976, have imposed significant new restrictions on the matching of records belonging to more than one Federal agency and on the matching of Federal agency records with those of other organizations. As a result of these new laws, and the climate of opinion in which they were developed, some agencies have limited access to their records for statistical purposes to an even greater extent than seems legally required. While the Subcommittee believes that some restrictions are essential to prevent the improper use of individual records, it also believes that some of the restrictions now in force have unduly inhibited the conduct of research studies based on exact matching of records. For example,

restrictions imposed by the Tax Reform Act have substantially increased the cost of follow-up studies to determine the mortality experience of persons exposed to potentially hazardous occupational or other environmental conditions. Formerly, IRS was able to screen lists of persons submitted by researchers and notify the re-32

searchers which persons had died, according to IRS records, and to provide information on state of residence and approximate time of death. The Tax Reform Act does not permit this use of IRS records, so researchers other than those in Federal agencies who are specifically granted access by the Act must now rely on other less complete and less centralized sources of information.

B. Recommendations

General

a. When Should Matching Be Used

When matching for statistical or research purposes is being considered, it is useful to assess whether matching is the best method of achieving the purpose. In some cases, the direct collection of data or some imputation technique, for example, might be better, As a minimum, the following factors should be considered in choosing the best method, giving each factor the appropriate weight for a specific application:

amount of error in the results resource cost time required

confidentiality and privacy considerations response burden

b. Choice between Exact and Statistical Matching If the conditions are such that there is a choice between exact and statistical matching, the factors listed above should be considered in choosing between the two types of matching. Great uncertainty exists regarding the error present in statistical match results; few attempts have been made to measure that error. Much more is known about the error present in exact match results. Taking into account the work that has been done and based upon theoretical considerations, in general the results of an exact match

are likely to contain far less error. No general comparison of resource costs and time required by exact and statistical matching can be made since these factors are very sensitive to the data files and methods used. Confidentiality and privacy considerations favor statistical matching, although the risk of disclosure from an exact match carried out for statistical purposes and done with the proper safeguards is small. When there is a choice between exact and statistical matching, the Subcommittee believes that a careful review of these factors would usually lead to the use of exact matching.

c. Documentation of Matches In cases in which the matched files will be used

by outsiders or when the matching techniques are of interest to outsiders, the matching should be documented carefully, even though substantial resources might be required for that task. Many of the matches which have been carried out have not been documented adequately. The documentation should in-clude descriptions of the files matched and the match-ing procedure. Adequate documentation allows others to assess the quality and usefulness of the

match and provides the information necessary for performing similar matches. Information about the cost of the match should be included. In addition, it is very important to compile and provide information concerning errors in the matched results. Documentation is especially important when the match is likely to be repeated or the results will be used for important policy decisions.

d. Public Release of Matched Data

If there is a demand for a matched microdata file, the release to the public of that file, after it has been determined that safeguards against inadvertent disclosure are adequate, should be encouraged. (The report of the Subcommittee on Disclosure-Avoidance Techniques of this Committee, Statistical Policy Working Paper 2, provides a detailed discussion of the disclosure problems which might be involved.) Even if the files which were matched were each previously reviewed for disclosure potential, another review is needed before the merged file can be released because the presence of more data for each unit (e.g., person) might make it easier to identify some units. Full use of matched data should be encouraged. Such matched

files frequently are of great use to researchers outside the group making the match.

e. Confidentiality Restrictions on Matching

Since exact matching is the only feasible or efficient method for many important statistical applications, the Subcommittee urges caution in the development and implementation of statutes, regulations and policies embodying confidentiality restrictions. In adopting measures for the protection of confidentiality, the distinction between record matching for administrative and for statistical purposes needs to be recognized. The purpose of administrative matching is to gather the information needed for taking administrative action with respect to each individual, and the individual's identification is therefore a key element of the matched file. In matching for statistical purposes the individual is of interest only as a link for bringing together relevant information; once that is done, the personal identifiers (name, etc.) are usually of no further use and are dropped from the

file, and the records become anonymous statistical units to be grouped with others for analysis. Interagency transfer of data with identifiers for this limited but important purpose should be recognized as a needed research tool and should be facilitated under strict controls protecting the files from unauthorized disclosure at any stage. Legislation permitting transfer of identifiable data for statistical purposes within protected enclaves" as recommended by the Director of the Office of Federal Statistical Policy and Standards (OFSPS, 1978) and by the Federal Statistical System Reorganization Project (1978) would, in the Subcommittee's judgment, be the most straightforward and effective means of achieving this goal.

2. Research

a. Exact Matching

More research on errors present in exact match results is needed. Research to develop improved methods of carrying out exact matches (e.g., assessing and reducing errors in various types of personal identifiers; better methods of determining optimal weights and thresholds) would be very useful.

b. Statistical Matching

A substantial amount of research on statistical matching is needed, regarding both optimal methods of matching and estimation of errors present in the matched results. Several promising research strategies have been suggested. For example, the results of exact and statistical matching of the same files can be compared. Also, tests to study the sensitivity of the results to the assumptions made in carrying out a match should be used more often.

APPENDIX I

Economics, Statistics, and Cooperatives Service

Example of Exact Matching

In the following, Section A describes exact matching approaches being developed for the purpose of unduplicating files, by the Economics, Statistics, and Cooperatives Service (ESCS), USDA as well as a more general examination of related topics. These topics are file considerations, match characteristic standardization, comparison pair reduction, and match rule selection. Section B describes the advantages and procedures for each selected match rule, while Section C examines practical problems in match rule application. Section D is a listing of papers from the technical notes of the List Sampling Frame Section of ESCS.

A. Exact Matching Considerations

In any match procedure, the first influence upon match rule selection is the constraints imposed by available data files. Once the goals of the match process have been adequately defined, it is necessary to determine whether existing files are suitable for

attainment of those goals. For each data file identified the following criteria are evaluated:

Cg

1. Coverage of file
2. Available match characteristics
3. Source definition of match characteristics
4. Quality of data for match characteristics
5. Source of maintenance procedures

Coverage and maintenance are the dominant factors in determining the number of files necessary to reach the match process goals. The available match characteristics, their definition and quality, substantially dictate the type of model to employ. If an accurate, unique identifier exists, this may be the only required characteristic to successfully unduplicate the files. In the ESCS match problem (attempting to develop an unduplicated list of farms which is as complete as possible) the input source files can be any files containing individual farm operations. No control exists over the match characteristics present, nor is there any control over

the definition or quality of those characteristics. However, a choice might be made to exclude a possible input file if quality is too low. No common format or content can be assured. This lack of standardization requires an additional match characteristic standardization step before a match rule can be applied. The degree of standardization needed depends totally on the input files. In many cases, the only standardizing necessary is a simple reformat operation. In the ESCS problem, the reformat used places name and address information into a standard order and form. The reformatted name and address fields are interrogated by programs which identify errors through word use coding. After possible errors have been reviewed and the standard format is accepted, the match characteristics are now accessible for a matching rule. There often exist too many paired comparisons to afford the match procedure so comparison reduction procedures are necessary. One or more characteristics of the file are used to divide the file in small portions, usually called blocks. The match rule will then be applied to all records within blocks. Blocking may be applied to name, address or identification variables. It is important to reiterate that blocking is used only to reduce total cost. If a match rule can

be applied without any or with very little blocking it should be. For the ESCS match problem, a separate sampling frame is to be built for each state so the state forms a first order of blocking. A second level of blocking results from processing individual, corporation and partnership files separately. (Any unusual name formats which cannot be clearly identified as individual or partnership are processed as corporate.) In the ESCS match problem, a block size of 300 or less is desired for individual class records. Specified blocking factors for individual class records in order of

use are surname code, first name initial group and location code. Surname codes are determined through use of a modification of the New York State Information and Identification System (NYSIIS). The first name initial grouping places together initials for which given names and common nicknames with different initials exist (such

as Bob, Robert; Bill, William; Dick, Richard). If surname code and first name initial group do not reduce a particular block to less than 300 records, the block is split into four quadrants based on longitude and latitude. Each record carries the latitude and longitude for its place name (city or town). If any resultant block is still too large, that quarter of state is again divided into four quadrants. For partnership records, the first two alphabetic surname codes are used for blocking. By definition, each partnership record must have at least two partners. Thus, Smith Bros would have the surname code for Smith twice as its blocking code. A partnership of Smith, Smith and Taylor would be found in the same block since only the first two alphabetic codes are used. Because of this "double blocking", no secondary level of blocking has been needed. For corporate records, the first stage of blocking is the corporate keynote with location used as a second stage when needed. A maximum block size of 500 is used for corporate records. The surname code divides most individual class records into acceptable sized blocks. For most states which have been run, about 99 percent of all final individual class blocks are created based on surname code only. One important feature of the ESCS match procedures is the ability to

match across blocks of records and across classes of records if records contain identifiers (box numbers, street addresses, telephone numbers, etc.) These procedures allow ESCS to detect nearly all of the duplication which was missed because of blocking while keeping costs to a fraction of making all possible match comparisons. Having completed these preliminary considerations, match rule selection is made. This is a most crucial step but the importance of correct selection is often not understood by users. Theoretical complexity and completeness does not necessarily mean best. Each particular alternative must be examined, weighing file structure and match characteristics before a reasonable selection is made. In considering match rules, we will again examine the ESCS procedures.

One type of match rule is intuitive in nature. Often this type of procedure stems from very reliable or unique match characteristics. Given either case one can use a very simple match rule, and accomplish about all that is necessary. This is true for the ESCS in blocking partnership records. A partnership record contains two or more surnames which have been alphabetized and coded in the data standardization procedure. A new variable consisting of the coded first two partner surnames is used as the major blocking

variable. This variable is nearly unique for partnerships with dissimilar surnames and yields small groups of partnerships with identical surnames. Newcombe, Kennedy, Axford, and James (1959) found a similar relationship when matching birth records using father's name and mother's maiden name.

A second type of match rule is empirical in nature. In using such a rule, more weight is given to current match characteristics in determining the proper criteria for a match rule. There usually exists some criterion for the match that is adaptable to match characteristic variations. In the ESCS development, such a procedure is employed to determine the proper threshold values for the individual class mathematical model. This procedure is described in Section B.

The final type of match rule is based upon some mathematical theory. Usually such a procedure is quite sensitive to file and match characteristic variations. These procedures are often developed to extract as much match information as possible from match characteristics of poor quality or completeness. In the ESCS case such a situation occurred with individual type records. The Fellegi-Sunter (1969) linkage technique was extensively modified to develop

a mathematical model which performs well over a wide range of file or match characteristic variations.

In most applications of a match rule, some questionable duplication is identified. If the matching results are to be improved these possible duplications must be examined and validated. However, the investigation should not stop there. To adequately evaluate a match rule, examples of the unquestioned decisions must also be examined. This later validation in both the matched and unmatched space is often left undone. If it were completed, many people would soon see the dilemma of exact matching. Any match rule only leads to guesses as to the true nature of duplication. These guesses are at best "mostly correct". With the present state of the data and the cost of match procedures, it is doubtful that significant increases in accuracy will be realized for the next several years.

B. Selected Match Rules

As the preceding discussion indicates, in the data preparation phase of the ESCS application, records are identified and separated according to three classes: individual, partnership, and corporate. Different matching techniques are employed to identify the duplication within each of these classes. Match rules have been chosen to fit the nature of the data available in records in each of these classes. The following briefly describes the procedures employed for partnership and individual classes. The corporate procedure parallels the partnership procedure.

1. Partnership Class

All partnership records have at least two surnames (not necessarily distinct). Given the discriminating power that two surnames afford and given generally the presence of additional match characteristics for these records, a simple set of decision rules is used to match records in this class. Thus, this match procedure is of the intuitive type, based on a set of predetermined rules which are applied uniformly to all records. A general outline of these rules follows, in the order in which they are tested. Comparison of

records takes place only within blocks of records for which the first two alphabetically ordered surnames receive the same surname code. Following the automated match process, as is true for all three classes, a manual review of these decisions takes place. Manual override capability is built into the system. The following steps illustrate the automated within block matching logic used. The process stops as soon as the first "if" statement is satisfied for a comparison pair.

a. If employer identification numbers are present and equal, the records are classified as links.

b. If the number of partners is not equal, the records are classified as non-links.

c. If _partnership keywords (e.g., Bros, Son) are present and not equal, the records are classified as non-links.

d. If first name initials are present and all equal, the records are classified as links.

e. If the distance between place names is greater than a parameter value, the records are declared non-links.

f. If box number or house number or both are present and equal, the records are classified as links.

g. Otherwise, the records are classified as possible links.

Logic used for corporate records is similar.

2. Individual Class Empirical Determinations

Even though a mathematical model has been established for matching individual class records, the same parameter values cannot be used for all applications (states in the ESCS case). Files in various applications differ in completeness of data available for linkage. For example, specific address information (street and house number or box number) is an important variable for linking individuals within a block. Matching address information receives a high agreement weight and nonmatching addresses receive considerable disagreement weight. All pairs of records which have a net agreement weight (total agreement weight for matching variables less the total disagreement weight) above a certain point or upper threshold will be called links. All below a lower threshold will be called non-links. All pairs of records between the two thresholds will be called probable links and must be manually reviewed. Setting a very low lower threshold will reduce the probability of false nonmatches but will also increase the amount of manual work required. Therefore, a

sampling procedure is used to set the desired threshold values for each application. The linkage model is first run with a lower threshold value such that all "true" duplicates would be expected to be linked together. A sample of linkage groups of various sizes created by this lower threshold value is selected to provide a cross section of the full file. All comparison pairs are outputted along with their corresponding weights. Each pair is resolved as a match or nonmatch. The empirical procedure then involves counts of number of comparison pairs that would be split apart by each incremental raising of the threshold along with counts of the number of these comparison pairs which actually represented the same individual. The sample counts are expanded to a total file basis so that the amount of duplication (false nonmatches) introduced by raising the threshold can be estimated along with the number of resolution decisions which will be left to make. In this ESCS name-matching example, reliability is expressed in terms of duplication left in the final master file. Records that are matched incorrectly will almost always be in the "probable link" category and will be resolved by manual procedures, so duplication is a bigger concern than percent of matches made correctly. Duplication occurs when the same individual is present in more than one input record and the matching procedures do not tie the

related records

together. In some of the first states in the ESCS project thresholds were set to allow .4 to .9 percent "new" duplication in various states. The reduction in manual workload exceeded 10 percent (as opposed to the workload necessary to achieve a "zero" percent duplication level) in every state, with reductions of manual workload as much as 30 percent for one state and 40 percent for another. Manual resolution of a sample may also have other benefits in terms of providing more information about the matching applications. In the ESCS example, individuals within blocks are sometimes observed manually who appear to be possible duplicates but have not been tied together by the computer models. This may occur if an individual is included in various files with different cities listed since city

discrepancy carries a high disagreement weight. The manual resolution procedures used by ESCS encourage reviewers to check these situations for duplication. Experience in three states has shown .65, .65 and 1.23 percent "original" duplication present in the master file that would be created. In the ESCS example subsequent procedures enable matching of records across blocks and across classes (individual, partnership and corporate) if they have identical addresses or matching identifiers such as telephone number. These procedures should eliminate much of the duplication left in by the thresholding decisions. Thus, the duplication percents obtained by the sampling procedures are maximums.

3. Individual Class Mathematical Model To extract the most information from a limited amount of data and to take into account the quality of the data on the file a mathematical model is used as the basis for the matching procedure for individual class records. The model is based on techniques suggested by Fellegi and Sunter (1 969). Briefly described, the space of all comparison pairs is divided into two disjoint sets: M = set of pairs representing the same individuals and U = set of pairs

representing different individuals. The outcome of each comparison pair can be represented by a vector of values representing the outcome of the comparison of each match characteristic. For each pair two probabilities are estimated:

1 . m prob. of observed outcome given pair is from M

2. u prob. of observed outcome given pair is from U These are converted into a test statistic or weight by:

weight = log10(m/u)

In the ESCS match problem, a separate list is to be built for each state. A frequency distribution is run on each name and address component in the reformatted individual files. The linkage models are based on this frequency distribution of the components so that agreement weights may differ for each value of the component (such as each surname). Since this frequency is run for each state individually, the agreement weights may also vary from state to state. For example, agreement on a surname such as Borowski may receive an agreement weight of 4.6 in South Carolina, but a weight of 2.6 in Wisconsin. The "error probabilities" above are set for 15 different components (prefix, given name, middle name, surname,

etc.). After review and manual resolution of a sample, the error probabilities for any or all components within state can be adjusted for production runs. Two threshold values are calculated to which the comparison pair weights are compared and classified as non-links (weight less than lower threshold), possible links (weight between thresholds), or links (weight greater than upper threshold). The final weight is a composite of agreement and disagreement weights for each item of linkage information. The following are several hypothetical examples of records being compared and the weights these comparisons might receive.

Example 1	Rec. I	Henry	P.	Agree	Rt 3	Lewisville
	Rec. 2	Henry		AgreeRt		
	3	Lewisville		Weights		
	+2.5	0		+ 4.1	+ 1.2+	
	2.1					

Total Weight = (+ 2.5) + 0 + (+ 4.1) + (+ 1.2) + (+2.1) = +9.9

Example 2

Rec. 1 William Bud Casey Rt l Box 87 Wheaton Rec. 2 Bill R

Casey Box 87 Wheaton Weights + 0.7 - 2.3 + 3.80+

2.6 + 2.5

Total Weight + 0.7) + 2.3) + (+ 3.8) + 0 +

+ 2.6) + (+ 2.5) = 7.3

Example 3 Rec. I Ed R. Johnston Rt 2 Lewisville Rec. 2

E R Johnston Rt 4 Wheaton

Weights + 1.9 + 1.2 + 3.6-

1.2

Total Weight = (+ 1.9) + (1.2) + (+ 3.6) +

(-1.2) = +5.5

Example 4 Rec. 1 George Smith Rt I Turkey Flats

Rec. 2 Richard Smith Rt I Turkey Flats Weights - 3.4+

1.8 + 0.3 + 3.1

Total Weight = (- 3.4) + (+ 1.8) + (+ 0.3) +

(+3.1) = +1.8

The models give much higher agreement for uncommon events than common ones (e.g., weight = 4.1 for the name Agree but only 1.8 for Smith). Data present for one record versus data missing for another record is not considered disagreement so no weight goes into the model for these cases. Route I is much more common than Route 3 so the agreement weights reflect this fact. If two place names (towns) differ, further address information is bypassed. The disagreement weights for place names are based upon their physical proximity. Adjacent towns would have a very low disagreement weight. There have been a number of modifications and extensions made to the theory in its application. Topics include weight calculation for surname code (which takes into account that surname code is the primary blocking factor), weight calculation for place name (in which distance is included as a variable), and weight calculation for social security

number (which illustrates a technique for using identifier numbers and for partitioning disagreement for these numbers). A listing of titles and authors of papers which are part of the technical notes of the List Sampling Frame Section of ESCS are included in Section D of this Appendix.

C. Practical Problems

The practical problems associated with using the above procedures are not uncommon to those using any procedure of the general type presented. An intuitive procedure, such as employed for partnership and corporate records, limits the user in that the rules are fixed and do not change with the file. While this is an advantage in that applying the procedure is a simple matter and does not change from one time to another, it does necessitate at least some manual followup to verify the results. The procedure is likely to be useful only when the match characteristics used are highly discriminatory and accurate. A model such as that used for individuals is more sensitive to the nature of the input files. While this can result in more reliable match results, it does require more effort on the user's part. To employ any such model, estimates of certain

parameters, such as error rates or cost functions, must be made prior to each application. The match results will be accurate only if these estimates are accurate. Empirical procedures are used by ESCS to establish accurate thresholds and error rates. Models of this type also depend on certain underlying assumptions about the data in order to apply these estimates in a linkage procedure. If these assumptions are violated then the applicability of the model becomes suspect. The examples of procedures presented above checked for matches only within blocks within class (individual, partnership and corporate) of record. A particular record could be represented in more than one class if both an individual and a firm name are used or could be represented in more than one block within a class if different names are sometimes used. Special procedures have been developed by ESCS which allow linkage across blocks and classes based on unique identifiers such as address, telephone number, employer's identification number, etc. A special feature of this supplemental matching involves the "generation" of trial records from secondary names associated with records of any class and from primary names in partnership class records to match against the individual class file.

D. Technical Papers

The following papers summarize research and modifications of matching theory completed during the development of the matching techniques for ESCS. Results of the papers below have been incorporated into the system for matching individual class names. Other areas of possible improvement have been identified and continue to be researched. These references arc not included in the bibliography.

1. Application of the Fellegi-Sunter Record Linkage Model to

Agricultural List FilesMax G. Arellano, 1976.

2. Weight Calculation for the Given Name Comparison-Max G. Arellano and Richard W. Coulter, 1976.

3. Weight Calculation for the Middle Name Comparison-Max G. Arellano, 1976.

4. Weight Calculation for the Surname Cormparison-Max G. Arellano and Richard W. Coulter, 1976.

5. Weight Calculation for the Place Name Cormparison-Max G. Arellano, 1976.

6. Processing of Comparison Pairs in Which Place Names Disagree-Richard W. Coulter, 1976.

7. Calculation of Weights for Partitioned Variable Comparisons-Max G. Arellano, 1976.

S. A Weight for "Junior" vs. Missing-Richard W. Coulter, 1976.

9. The Estimation of Component Error Probabilities for Record Linkage PurposesMax G. Arellano, 1975.

39

10. The Estimation of P(M)-MaxG.Arellano, 1975. 1976.

11. Sampling Size in Estimating Component Number for Matching Purposes-Max G. Error Probabilities-Ricbard W. Coulter, Arellano, 1976.

12. Optimum Utilization of the Social Security

APPENDIX 11

Office of Research and Statistics Example of Statistical Matching

A. Introduction and Input Files

The 1972 ORS Statistical Match File was constructed in the Office of Research and Statistics, Social Security Administration, by statistically matching the 1973 Exact Match (EM) file and the Augmentation File (AF). The Statistical Match File is being used to examine the role of social security in the tax-transfer system. In order to carry out that research, more tax return data than are contained in the EM were needed. Particularly important was the addition of amounts of individual Federal income tax liabilities, which are not contained in the EM. That necessary information was added in this statistical match. The version of the EM used contained the following data sources:

1. March 1973 Current Population Survey (CPS) (demographic, work experience, income, and family composition data)
2. Social Security Administration (SSA) Summary Earnings Record (SER) extract (earnings and demographic data)
3. Internal Revenue Service (IRS) Individual Master Tax File (IMF) extract for 1972 (limited income data)

As its name suggests, the EM was the product of an exact match, primarily using social security numbers, among those three data sources.

The AF contained the following data sources:

1. SSA SER extract (earnings and demographic data)
2. IRS Statistics of Income (SOI) subsample for 1972 (detailed income and tax data)

The AF was the result of an exact match, using social security numbers, between those two data sources.

B. Matching Method

In this statistical match, for each unit in the base file (the EM),

the nonbase file (the AF) was searched for the observation which "most closely resembled" what the exact match data for that EM record were thought to be. That is, for each EM record which contained income tax return data, the AF was searched for the observation which was thought to most closely resemble the tax return actually filed by that unit, and that unit's SSA data. In this match, there were several variables which were defined (almost) identically in the two files and which were obtained from the same data source. (The AF was constructed with this comparability in mind.) For those variables, the AF values searched for would be identical to (or very close to) the EM values, and those searched for values could be determined with accuracy. This match was made by separating both files into comparable cell categories and using ranges and a distance function to choose, for each EM record, the best match within the cell and ranges. The variables used to make the match are shown in Table 1. The first 14 variables can be considered to be common to the two files-that is, they have the same (or very nearly the same) definition and can be expected to have the same (or very nearly the same) error pattern in the two files. In other words, in an exact match carried out without error, values for the pair in the two files would be identical (or very nearly the same). The first ten variables

in Table I were used as cell classifiers (see Table 2). Age and adjusted gross income (AGI) were used as ranges around the EM value. The age range was the EM value plus or minus five years. For most records, the AGI range was the EM value plus or minus ten percent, with a minimum range of $1,000 (see Table 3). Nineteen variables (all variables except number of taxpayers,

About 83 percent of the EM records had identical values for all 15 variables, and more than 99 percent had 12 or more fields equal. It should be noted that, in general, nonzero income amounts were not required to be equal in this test.

Another indicator is, using the pair of variables as the unit of observation, the percent of EM records in which the AF value is identical; these data are shown for several variables in Table 7.

These variables all had important roles in the matching and would be expected to have high percentages of identical values; for the most part that was true. However, these percentages should be interpreted with caution; they can vary widely among subgroups in the file. For example, returns with Schedule C in the EM had Schedule C in the AF in only about 84 percent of the cases. Some rematching is being done to improve the correspondence for some variables.

D. Tables

Table 1-Variables Used in the Statistical Match

	EM	AF	Source of Variable	Source of Data'	Variable of Data
1.	Number of Taxpayers'		IRS	IRS	
2.	Sex'	SSA	SSA		
3.	Race	SSA	SSA		
4.	Marital Status	IRS	IRS		
5.	Number of Dependent Exemptions	IRS	IRS		
6.	Type of Earnings	SSA	SSA		
7.	Size of Earnings	SSA	SSA		
8.	Wage and Salary Income	IRS	IRS		

9. Dividend Income (after exclusion)IRS IRS
10. Interest Income IRS IRS
11. Age SSA SSA
12. Adjusted Gross Income' IRS IRS
13. Net Adjusted Gross Income' IRS IRS
14. Number of Age and Blind Exemptions IRS IRS
15. Presence of Schedule C (nonfarm business income) IRS IRS
16. Presence of Schedule E (supplemental income)IRS IRS
17. Presence of Schedule D (capital gain or loss) IRS IRS
18. Presence of Schedule SE (self-employment income) IRS IRS
19. Presence of Schedule F (farm income) IRS IRS
20. Presence of Rent and/or Royalty Income CPS IRS
21. Presence of Pension Income CPS IRS
22. Home Ownership CPS IRS

a IRS = internal Revenue Service

SSA = Social Security Administration

CPS = Current Population Survey

bNot used in the distance function.

I Defined as adjusted gross income minus $750 times the total number

of exemptions.

APPENDIX III

Selected Examples of Exact Matching

Examples

A. Record Check Studies of Population Coverage

B. Matching of Probation Department and Census Records

C. Computer Linkage of Health and Vital Records: Death Clearance

D. Use of Census Matching for Study of Psychiatric Admission Rates

E. June 1975 Retired Uniformed Services Study

F. Federal Annuitants-Unemployment Compensation Benefits Study

G. Office of Education Income Validation Study

H. Department of Defense Study of Military Compensation

1. Department of the Treasury-Social Security Administration Match Study

J. G.I. Bill Training Study

K. 1973 Current Population Survey-Internal Revenue Service-Social Security Administration Exact Match Study

L. Statistics Canada Health Division Matching Applications

M. Statistics Canada Agriculture Division Matching Applications

A. Record Check Studies of Population Coverage

(Part of 1960 Population & Housing Census Evaluation & Research Program)

1. Data sets: a. 1960 Population Census enumera-tion records.

b. Samples from:

(I) 1950 census records: 3-stage sample: county (333 CPS sample areas)-Enumeration District (ED) (1067)-persons

(2,600); sampling rate 1: 60,000.

References which appear only in this appendix are not included in the Bibliography.

(2) Registered births after 4/1/50 and before 4/1/60: 2-stage sample: counties (same 333 CPS areas)-birth registrations (4,500); sampling rate 1:8,700.

(3) 1950 Post-Enumeration Survey (PES): persons detected by PES as missed in 1950 census: subsample of 273 persons; sampling rate I : 1 1,400 persons estimated as missed. Sample design = 1950 PES (multi-stage).

(4) Aliens residing in U.S. in Jan. 1960, registered with Immigration and Naturalization Service. Systematic sample of individuals in I I states with 80 percent of registered aliens; Systematic sample of individuals in 5-state sample drawn from other states.

Total: 209 persons; sampling rate 1: 14,000.

2. Purpose: Census coverage evaluation.

3. Type of match, and procedure: Exact match, longitudinal reverse record check, manual.

a. Determination of sample persons' April 1960 address, by mail (starting with a post office check), and personal interview if no reply.

b. Search of 1960 Census records: spot 1960 address on map, determine ED, locate address in census records.

c. Clerical coding of degree of match, based on name and address and on supplemental information.

d. Field reconciliation of unmatched and doubtful cases, by letter, phone, visit.

Address-name codes were assigned according to whether the address and the names were identical, similar, or non-contradictory; the 3 terms were defined specifically and separately for addresses and for names. Supplemental information codes were assigned on the basis of the coder's interpretation of the additional evidence available in each case; the 5 categories of this code could not be defined

specifically like those of the address-name code, but

the categories were illustrated by a number of annotated examples used for the training of the coders. Independent verification showed a very low error rate in the address-name codes and a high degree of consistency in the supplemental code. On the basis of the combination of the two codes, each case was classified as "Matched" (i.e. with clear evidence that the person was enumerated in the census) or "Nonmatched" (apparently missed in the census, or doubtful). Nonmatched cases were reviewed and subjected to field reconciliation and an additional census search. Emphasis was placed on minimizing erroneous nonmatches and net matching error.

4. Publications: a. Overall report with results: Record Check

Studies of Population Coverage. Series ER 60

No. 2, Bureau of the Census, 1964.

b. Detailed description of the matching procedure, codes and definitions, matching rules, with illustrative examples: "Matching for Census Coverage Checks," by Walter M. Perkins and Charles D. Jones. In: 1965 Proceedings of the American Statistical Association, Social Statistics Section, pp. 122-141.

5. Contacts: Charles D. Jones, Chief, Statistical Methods Division, Bureau of the Census, (principal author of 4.a and coauthor of 4.b); or Hans J. Muller, Statistical Methods Division, Bureau of the Census.

B. Matching of Probation Department and Census Records

(Southern California Records Matching Project)

Initiated in 1963

Research supported by a National Institute of Mental Health

(NIMH) grant.

1. Data sets: a. 1960 Population census records

b. All (13,315) "official" cases of juvenile delinquents age 10-17 referred to Los Angeles County Probation Department, 7/l/59-12/ 31/60 (18 months centered on census date). (4/5 male; 2/3 Anglo, 1/6 Negro, 1/6 Spanish surname)

2. Purpose: Improved delinquency rate information (as compared to rates based on aggregate data from the 2 systems) including characteristics of household.

3. Type of match and procedure: Exact matching, by computer with a final visual step (1/6 also matched manually all the way).

a. Data extracted from case files: "Intake" data on age, sex, race, offense, address of adult family member likely to be in Census; allocation to census tract.

b. "Feasibility phase": 2,316 cases (1/6) matched visually to census records at Jeffersonville. Traced cases allocated

to ED's; files searched for records of juveniles and/or adults in household; if failure on 1st address, search at other given address(es). Institutions included. Key criteria: age, sex, race, relationship to head. If not a "complete match" (Juvenile + household head) cases are returned for obtaining additional addresses. Juvenile + adult located in 84 percent of cases.

c. Before the rest (5/6) could be matched, the census data were transferred to microfilm; the original records were destroyed. Use of the microfilm reels would have required a prohibitive increase in time and money.

Alternative: use of computer tapes of 25 percent census sample, proved almost as effective as (b) with a substantial reduction in cost; besides, the 25 percent sample is more useful because it includes data on more variables.

The probation data were matched to the census listing books which show for each household: address, surname of household head, number of persons, sample status, FOSDIC page number, ED number; 23.3 percent were found to correspond to the 25 percent

sample. For these the ED and FOSDIC page number, age, sex, race of the juvenile and relation to head were punched on cards. The cards were matched to a 25 percent census tape for L.A. County specifically prepared for this project. Match failures were handled "similarly" to (b), but there were differences. As final step, the 25 percent microfilm records were used to verify unmatched and marginal cases. (The 25 percent procedure seems to have been used for all cases, including the ones already put through the visual search.)

d. As each case was located, a tape of the delinquent population was created, with the census data on population and housing characteristics of the household. A general population tape file, with data for all families and housing units with one or more children 10-17 (delinquency rate denominators) was derived from the original L.A.

county tape prepared for (c).

Results: Apparently almost all case addresses were found in the listing books, and only 47 cases are shown as "status undetermined". However, a sub-stantial proportion of the persons-juvenile and reference adult-were not found in the actual matching.

Rates	Percent Found Cases Searched	Neither Juvenile & Adult	Matching Juvenile Only	Total Adult Only	one found, %
Visual (Feasibility Study) 1	2,125	84.0	2.7	3.8	9.5
Computer (25% Census Sample)	2,919	77.8	2.0	6.7	13.5

Higher rates from visual matching may be due to more elaborate

efforts to obtain a match: for the feasibility sample, initially unmatched cases were followed up by reviewing school, public assistance, vital statistics, Youth authority, and Juvenile Index records and by a detailed study of Probation Department case folders, in an attempt to get additional addresses; for the unmatched cases from the computer match, only the probation files were examined. Results from the feasibility study have been tabulated. No significant variations between matched and unmatched cases were found with respect to sex or race; differences by offense were not statistically significant, although there was a tendency to achieve more success in locating juveniles processed for auto theft, major traffic and property violations than for sex delinquencies and offenses against the person (robbery, rape, etc.). The findings suggest that the matched cases are representative of the Probation Department universe of "official" cases.

- 96 percent agreement between probation and census entries for sex, race, relationship, and over 92 percent forage (- I year).

- No major attrition over the 18 months time span (range of match rates: 73.5 percent (July 59, I st study month)- 94.3 percent (Dec. 59); overall

rate 86.7 percent; only 3 months under 80 percent). Estimates of

the extent to which underenumeration and sampling errors affected the ratio actually obtained, have not been developed yet.

4. Publication-The Matching of Census and Probation Department Record Systems, John E. Simpson and Maurice Van Arnold, Jr. (U. of Southern California). In: 1965 Proceedings of the American Statistical Association, Social Statistics Section, pp. 1 1 6-1 2 1.

5. Contacts:

a. Simpson and Van Arnold

b. Dr. George Sabagh U. of Southern Calif., Dept. of Sociology and Anthropology, Pop. Research Lab.; Youth Studies Center.

c. Census Bureau coordinator: John C. Beresford, Population Division.

C. Computer Linkage of Health and Vital Records: Death Clearance

New York City Department of Health, under contract with the

National Center for Health Statistics (NCHS)

1. Data sets:

a. Death file: magnetic tapes made from routine death index punchcards of N.Y.C. Health Department, including all deaths occurring in the city and deaths of city residents reported to N.Y.C. as occurred outside the city; 1961-63. Size of file: 281,208.

b. Coronary heart disease (CHD) population of HIP (Health Insurance Plan of Greater New York): 176,481 members of medical groups in the HIP-CHD study for 1961, 1962 and 1963. Records: HIP enrollment cards.

2. Purpose: To study the feasibility of large-scale computer linkage when only limited amounts of identifying information are available.

3. Type of match: Exact match, by computer, with a final clerical step.

I st step: Soundex coding of names (first and

last)

2nd step: The computer program brings together records from the 2

files having the same Soundex codes, and compares the HIP-death pairs

to see whether there is agreement on common items of infor-

mation: surname, first name, age.

[The HIP enrollment card does not show race; sex is not a very

discriminating item; the two records have no other useful information

in common.] The program produces a listing of the pairs that meet a

set of minimum matching criteria: Exact agreement on Soundex code of

first and last names, and age agreement within 5 years. (Many records

appear in more than one pair.)

3rd step: Clerical elimination of pairs that do not seem to constitute valid linkages; and validation of remaining pairs by using information that could not be used in the computer program but can be obtained from other HIP records.

This includes verifying, through HIP contact with their members, a list of deaths found by the computer procedure but not previously known to HIP, and obtaining lists of deaths known to HIP but not detected by the computer procedure. (Since the two procedures for finding deaths are independent, this may make it possible to estimate the number of deaths missed by both.)

Fiiidiiigs: The computer run reduced the ai)proximately 176,000 medical records and 281,000 death records to 89,306 possible matched pairs with exact agreement on the Soundex codes of first and last names and age agreement within +/- 5 years. This includes:

7,036 pairs with exact agreement on first and last names, and age +/- I year;

13,835 pairs with exact name agreement but age differences of 1 to 5 years; 13,424 pairs with age agreement years, and exact

agreement on first or last name; 5,615 pairs with age agreement +/-

I year,

and no exact agreement on either name;

34,970 pairs with age difference 1-5 years, agreement on first or

last name;

14,426 pairs with age difference 1-5 years, no exact agreement on

either name.

(Findings from the clerical review are not reported

in the source.)

Conclusions: It is very unlikely that many-if any-true matches

would not be included in the group of 89,000. Each of the subgroups

listed above

probably includes some true matches, in decreasing proportions of

each subgroup. Most of the true matches should be among the 7036

pairs with exact name agreement and age within I year. However,

since the expected number of deaths in the patient group under study

is stated to be about 3000, even this subgroup must include a

substantial proportion of spurious matches. Clearly, name and age

are not sufficiently discriminating in this population; additional

identification items are needed for selecting the true matches. On the other hand, the 7036 pairs with exact name agreement probably do not include all true matches because they do not allow for spelling variations. The Soundex code remedies this by allowing for such variations, but it also pulls in many pairs with what really are different names. This again emphasizes the need for additional matching information (which, in this case, could not be included in the computer program but had to be done through a clerical operation). Under the conditions of this study, the value of the combination of Soundex code and computer matching lies in the quick reduction of the mass of original data to a more manageable number of possible matched pairs that can then be investigated clerically. The investigators hoped that their results would enable them to modify Soundex to make it a finer, more efficient "noise filter" for names; nothing is said on whether any work in that direction was actually undertaken.

4. Publication: "The Methodology of Computer Linkage of Health and Vital Records." David M. Nitzberg (Harvard School of Public Health) and Hyman Sardy (Brooklyn College), In: 1965

Proceedings of the American Statistical Association, Social Statistics Section, pp. 100- 1 06.

5. Contacts: This study was undertaken by the N.Y. City Department of Health under a contract with NCHS, to develop computer death clearance techniques. Project Director was Dr. Paul M. Densen, Deputy Commissioner of Health, N.Y.C. Sidney Binder, Chief of the Data Processing Div. of NCHS, "assisted". The programs (for IBM 7010) were written by Dr. H. S. Levine (HIP) and J. Hayden.

(The same group has also worked on linking other record groups with the NYC death file; the HIP-CHD group is the largest one and the only one reported in the publication).

D. Use of Census Matching for Study of underenumeration)

and the degree of incom- Psychiatric Admission Rates

pleteness is not known in either case, for the specific categories involved

(i.e. in this case, NIMH study on persons admitted to all psychiatric census undercoverage estimates were available facilities in Maryland (14,450) and Louisiana for the U.S. but not for Maryland and Louisiana).

(13,036) during the year following the 1960 census. It was concluded

that under certain assumptions

1. Data sets: he ratio of the observed admission rates is a consistent estimate of the "true" relative risk.

a. 1960 Population Census records.

b. Institutional data, including supplementary

4.Publication: "Use of Census Matching for Study data collected for this study: name, sex, color, of Psychiatric Admission Rates." Earl S. Pollack birthdate, psychological diagnosis, facility (National Institute of Mental Health). In: 1965 where admitted, admissions history, residence tables. Proceedings of the American Statistical Association,- on admission and on 4/l/60, name of house- Social Statistics Section, pp. 107-115, 9 hold head on 4/l/60.

2. Purpose: To study the feasibility of determining E. June

1975 Retired Uniformed differential admission rates for specific

population groups (by sex, age, race, diagnosis, etc.)

3. Type of match, and procedure: exact match, manual.

The institutional data were posted on transcription sheets and

given to the Census Bureau for matching and tabulation. They

were coded with ED (census enumeration district) numbers, based

on the 4/60 addresses; where addresses were uncertain, one entry

could have several possible ED numbers. The data were then matched against the census ED books. "Not Matched" if not found in any of the possible ED's indicated.

Findings: Matched: 67% of Louisiana, 64% of Maryland patients.

Matching was most successful for: under 18; males; whites; household heads and close relatives; married. Matching least successful for: age 18-24; females; non-whites; alcoholics. Match rates tended to be high in categories where the census undercount tended to be low. Low match rates in some groups may be due to underenumeration in the census (alcoholics, etc.)

Possible reasons for failure to match (not investigated) :

(1) Inadequateaddresses

(2) Name and age differences

(3) Clerical error

(4) Persons not enumerated in census.

[A methodology was developed for evaluating the differences in

admission rates between 2 population categories when the numerators are incomplete (because for some admissions the matching census entries were not found) and the denominators are understated (because of census

Services Study

I . Data Sets: The Civil Service Commission Central Personnel Data File as of June 1975 was matched against tapes of retired uniformed services personnel receiving benefits from eight finance centers.

2. Purpose: The study was conducted to provide Congress indications of impact of reemployment of retired uniformed services personnel in the Federal civilian service.

3. Type of Match: An exact match on social security number was performed to produce outputs which describe Federal employees who are retired from the uniformed services. Data matched included date of retirement, length of service, uniformed service component, basis of retirement, military pay grade and retirement pay as well as approxi-mately 15 demographic characteristics from current

employment.

4. Reference: A report was prepared for the House Subcommittee on Manpower and Civil Service of the House Post Office and Civil Service Committee.

5. Contact: William Anderson, Office of Personnel Management.

F. Federal Annuitants-Unemployment Compensation Benefits Study

1. Data Sets: The Civil Service Commission (CSC) Federal Retiree File and Central Personnel Data

37 (a) 'ne ratio of the cross products of match rates and under-coverage rates in the 2 categories (mie2/m2ei) must be close to I (this condition applies if only the actually matched admissions are used as numerators in the admission rates); or

(b) the ratio of the census undercoverage rates in the 2 categories (e2/ex) must be close to I (if an estimate of missed matches is added to the matched admissions in the admission

rates).- 1

File current Status File were matched with Department of Labor Unemployment Compensation Benefits input files from 24 States.

2. Purpose: For the 24 States the study was to determine the incidence of Federal civilian employees receiving an annuity from the Federal Government and receiving unemployment insurance benefit payments concurrently. Reports were produced for 1974 and 1975 showing the number of Federal retirees and the number of those receiving unemployment benefits.

3. Type of Match: An exact match on social security numbers was used to link the Unemployment Compensation data on year of first payment and State with Civil Service data on year of retirement and State of last duty station.

5. Contact: Robert Penn, Office of Personnel Management.

G. Office of Education Income Validation Study

I . Data Sets: Applications for the Basic Educational Opportunity Grant (BEOG) program were matched against Internal Revenue Service (IRS) files for tax years 1973 and 1974.

2. Purpose: The match was performed to identify categories of applicants that most frequently report income, tax and dependent data which vary from IRS reported data.

3. Type of Match: An exact match using social security numbers was performed. Two Office of Education contractors were involved in the study. One contractor selected the sample and provided IRS with a tape of control numbers, social security numbers, and name control data. The second contractor was provided application data for the sample identified by control number only, with no social security numbers or names. After matching the social security numbers with reported tax data IRS provided the second contractor with a file of relevant tax report data identified

only by control number. Original data tapes were destroyed after the IRS matching was completed.

5. Contacts: Gloria Koteen or Paul E. Grayson, Department of the Treasury.

H. Department of Defense Study of Military Compensation

1. Data Sets: A Department of Defense sample of the military population was matched against Internal Revenue Service (IRS) files for tax year 1974.

2. Purpose: The Department of Defense wanted to develop information on the tax advantage of currently non-taxable allowances paid on military personnel.

3. Type of Match: An exact match using social security numbers was performed. The Department of Defense provided a sample file identified only by social security numbers and data cell (pay grade, length of service, Branch of Service). IRS matched this file against 1974 tax year records and

created a file with 10 data items for analysis. Outliers with extremely large incomes or with 10 or more exemptions were removed from the study.

5. Contacts: Gloria Koteen or Paul E. Grayson, Department of the Treasury.

1. Department of the Treasury-Social Security Administration Match Study

I . Data Sets: Department of the Treasury Statistics of Income individual income tax return and estate tax return files were matched against Social Security Administration Summary Earnings Record Files and Limit Special Beneficiary Files.

2. Purpose: The Office of Tax Analysis, Department of the Treasury, was interested in studying the effect of income taxes and estate taxes on earnings. The Social Security Administration (SSA) wished to add income (and wealth) items unavailable on SSA earnings records for use in policy simulation models of alternative taxtransfer systems.

3. Type of Match: An exact match using social security numbers

was used. SSA was provided social security numbers (SSN's) for all sampled returns and prepared data files from the Summary Earnings Record Files and Limit Special Beneficiary Files for these SSN'S. These data were then linked with statistical extracts from the tax returns.

5. Contacts: Nelson McClung and Jack Blacksin, Treasury, and Fritz Scheuren and Henry Patt, Social Security Administration.

J. G.I. Bill Training Study

1. Data Sets: Department of Defense data files on military enlistees separating from active duty in

1969 after completing one term of service were Specific objectives of SSA included studying matched with Veterans' Administration program effects of alternative ways of determining social participation information files and Social Security security benefits, summarizing lifetime covered Administration files of earnings information. earnings patterns of persons contributing to 2.Purpose: Linkage was performed to determine social security, obtaining some additional infor mation about noncovered earnings, examination training programs on future earnings. effects of participation in G.I. Bill Benefits paid Another of age reporting differences among matched item of study was assessment of effects of the sources, and use in policy simulation models of Armed Forces Qualifying Test waiver on service the tax-transfer system. and post-service earnings.Objectives of other participants included assist-3.Type of Match: The files were matched through ing in construction of a "corrected" income size an exact match using social security number.distribution of the U.S. population, examining Samples of G.I. Bill users and non-users were differences in income reporting in an attempt to selected for which Department of Defense and "improve" the CPS interview schedule, and con-Veterans' Administration data were matched.stitution of control groups for research into man-These matched files were transferred to the Social power training programs.

Security Administration for matching and addition of earnings data. All identifiers were then removed before analysis was performed.

5. Contact: Dave O'Neill, Sue Ross, The Public Research Institute (Division of Center for Naval Analysis), 1401 Wilson Blvd, Arlington, VA 22209; Wendy Alvey, Fritz Scheuren, Office of Research and Statistics, Social Security Administration.

K. 1973 Current Population SurveyInternal Revenue Service-Social Security Administration Exact Match Study

I . Data Sets: Current Population Survey (CPS) control card data, basic CPS information and March (and June) supplement items for persons interviewed in the March 1973 CPS were matched with the Internal Revenue Service (IRS) 1972 Individual Income Tax Master File and the Social Security Administration (SSA) Summary Earnings Record File, Quarterly Wage File, Benefits in Force File, Limit Special File, Master Beneficiary Record File and National Employee Index File.

2. Purpose: The study had several objectives. The overall objectives were evaluation and "correction" of income data from matched sources, exploration of weighting and control procedures used to adjust for non-interviews and survey

undercoverage, augmentation of survey data with information missing because it was not asked or was not provided, and creation of a public-use file available to statisticians and researchers both within and outside the Federal government.

3. Type of Match: An exact match procedure using social security number was employed, but confirmatory variables such as name, race, sex and date of birth were also examined. The confirmatory variables were also used in searching for missing account numbers. Census provided tapes to SSA of control card and CPS data as well as abstracted IRS income tax information. SSA did all of the other matching in stages. Throughout the matching procedure, weighting factors were introduced to "correct" for undercoverage, mismatching and erroneous mismatching.

4. References: "The 1973 CPS-IRS-SSA Exact Match Study: Past, Present, and Future," by Beth Kilss and Fritz Scheuren, a paper presented at the NBER Workshop on Policy Analysis with Social Security Research Files, Williamsburg,

Virginia, March 16, 1978 (in Policy Analysis with Social Securitv Research Files, the proceedings volume); "Exact Match Research Using the March 1973 Current Population Survey Initial States," by Frederick J. Scheuren et al., Studies from Ititeragency Data Liiikages, No. 4, Office of Research and Statistics, Social Security Administration, July 1975; other reports in the Studies from Interagency Data Liiikages series.

5. Contact: Roger Herriot and Emmett Spiers, Census Bureau; William Smith and Peter Sailer, Internal Revenue Service; Fritz Scheuren and Beth Kilss, Social Security Administration.

L. Statistics Canada Health Division

Matching Applications

1. Data Sets: Statistics Canada has used matching techniques for several Health Division Studies

involving medium and large-sized files. The data sets matched have included:

a. Admission/Separation records for TB patients during the period 1951-1960 matched with mortality data for 1951-1973 and cancer incidence data for 1968-1973.

b. A sample of occupational records for selected industries for the period 1965-1971 matched with mortality data for 1965-1973.

c. A file of uranium miners for the period 1955-1974 matched against mortality data for the period 1955-1974.

d. A file of infant deaths following births in 1971 matched with relevant birth records.

c. All known records for death due to anencephaly, a birth defect, in 1969-1972 were linked to birth records for those

years.

f. A file of birth anomalies occurring in 1971 matched against files of birth and stillbirth records for 1971.

2. Purpose: Each of these studies was performed to study relationships between environment or heredity factors and birth defects or illnesses and deaths. The study of former TB patients was made to determine if the drug INH used for some TB patients is a potential carcinogen. The Occupational Record Study and the Uranium Miner Study were made to investigate relationships between occupations and potential causes of cancer and death, particularly the relationships when exposed to uranium dust. The Birth Record studies were performed to evaluate potential relationships between birth characteristics and birth defects or infant deaths.

3. Type of Match: These were exact matches which used all available name, address and demographic data. The INH and Occupational studies involved such large data files that

manual resolution was impractical. Also the objective of the studies was analysis of data obtained by matching records not the matching itself. Thus, initially, a high threshold was set and only matched pairs above the threshold were analyzed. The threshold was then progressively lowered and the analysis repeated an increasing number of matched records. By this means the sensitivity of the analysis to the matching procedure could be checked.

Some of the birth records could be matched on a unique registration number. The birth record matching included information for both father's name and mother's name. The congenital anomaly study was primarily an exact match study with several iterations. An anomaly file record was allowed to match several birth records so that the "best" match could be selected.

5. Contact: Elizabeth Coppack, Statistics Canada.

M. Statisfics Canada Agriculture Division Matching Applications

I . Data Sets: Statistics Canada h

as explored record linkage and matching techniques for a number of agricultural applications. Two specific matching efforts were:

a. Files of 1971 Census of Population variables were matched against 1971 Census of Agriculture files.

b. The Farm Register file was matched against 1976 Census of Agriculture files.

2. Purpose: The matching of the Censuses of Agriculture and Population was done to bring together variables from the two sources for publication of cross tabulations. The Farm RegisterCensus of Agriculture match was performed to update the Farm Register as a mailing list source and a source of up-to-date commodity data.

3. Type of Match: This was an exact match. The Census of Population contained a household number for identification but the integrity of this number was not consistently ensured for the Census of Agriculture. So additional matching based on farm operator age was necessary.

For the Farm Register-Census of Agriculture linkage, matching

used NYSIIS operator name within postal office as the minimum match criteria. In-house address decoding utilities were also used.

4. Reference: 1971 Statistics Canada Census Publications 96712 to 96717 contain the results of the Censuses of Population and Agriculture matching.

5. Contact: Censuses of Population and Agriculture match-Wilson G. Freeman; Farm Register Match-R.W. Freeman; both of Statistics Canada.

Bibliography

Note: (E) denotes that the reference is concerned with exact matching.

(S) denotes that the reference is concerned with statistical matching.

(E,S) denotes that the reference is concerned with both exact and statistical matching.

(This Bibliography excludes references which appear only in Appendices I or 111.)

Alter, Horst E. (1974). "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970." Annals of Economic and Social Measurement (April) 2: 373-394. (S) Althauser, Robert P., and Rubin, Donald (1969). "The Computerized Construction of a Matched Sample." American Journal of Sociology (Sepiember) 76: 325-46. (S) Alvey, Wendy, and Cobleigh, Cynthia (1976). "Exploration of Differences Between Linked Current Population Survey and Social Security Earnings Data for 1972." 1975 Proceedings of the American Statistical Association, Social Statistics Section, 121-28. (E) Armington, Catherine, and Odle, Marjorie (1975). "Creating the MERGE-70 File: Data Folding and Linking." Research on Microdata File, Based on Field

Surveys and Tax Returns. Working Paper 1. The Brookings Institution (June). Mimeographed. (S) Barr, Richard S., and Turner, J. Scott (1978 a). "A New, Linear Programming Approach to Microdata File Merging." In 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (Barr and Turner's reply to Goldman also appears in that Volume.) (S)

Barr, Richard S., and Turner, J. Scott (1978 b). "New Techniques for Statistical Merging of Microdata Files." Paper prepared for the Conference on Microeconomic Simulation Models for the Analysis of Public Policy, National Academy of Sciences, (March.) (S)

Barr, Richard S., and Turner, J. Scott (1979). "Microdata File Merging Through Large-Scale Network Technology." Working Paper 79-100, Edwin L. Cox School of Business, Southern Methodist University (May). (S)

Beebout, Harold; Doyle, Pat; and Kendall, Allen (1976). "Estimation of Food Stamp Participation and Cost for 1977: A Microsimulation Approach (Final Report)." MPR Working Paper #E-48, Mathematica Policy Research, Inc. (July). (S) Budd, Edward C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income." Review of Income and Wealth (December) 17: 317-33. (S) Budd, Edward

C. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 349-54. (S) Budd, Edward C., and Radner, Daniel B. (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964". American Economic Review (May) LIX: 435-49. (S) Budd, Edward C., and Radner, Daniel B. (1975). "The Bureau of Economic Analysis and Current Population Survey Size Distributions: Some Comparisons for 1964", in James D. Smith, ed., The Personal Distribution of Income and Wealth, Studies in Income and Wealth, 39: 449-558. (S) Budd, Edward C.; Radner, Daniel B.; and Hinrichs, John C. (1973). "Size Distribution of Family Personal Income: Methodology and Estimates for 1964." Bureau of Economic Analysis Staff Paper No. 21. U.S. Department of Commerce (June). (S) Bureau of the Census (1974). "Unimatch I Users Manual-A Record Linkage System." Census Use Study. Washington. (March) (E,S)

Colledge, M.J.; Johnson, J.H.; Pare, R.; and Sande, I.G. (1979). "Large Scale Imputation of Survey Data." 1978 Proceedings of the American Statistical Association, Survey Research Methods Secion, 431-6. (S)

Coulter, Richard W. (1977). "An Application of a Theory for Record Linkage." Paper presented at

the April 6 meeting of the Washington Statistical Society. Washington, D.C. (E) Duncan, Joseph W. (1976). "Confidentiality and the Future of the U.S. Statistical System." American Statistician (May). 30: 54-59. (E,S) Federal Statistical System Project (1978). Issues and Options (draft), Office of Management and Budget. (E,S) Fellegi, Ivan P. (1978). "Discussion." 1977 Proeedings of the American Statistical Association, Social Statistics Section, 762-4. (E,S) Fellegi, Ivan P., and Sunter, Alan B. (1969). "A Theory for Record Linkage." Journal of the American Statistical Association, 64: 1183-1210. (E) Goldman, Alan J. (1978). "Comment." in 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (S) Hollenbeck, Kevin (1978). "A Design for Creating a New CHRDS Data Base Using the Survey of Income and Education and Annual Housing Survey." Discussion Series Ref. #

7309-01-004, Mathematica Policy Research, Inc. (September 29). (S) Hollenbeck, Kevin, and Doyle, Pat (1979). "Distributional Characteristics of a Merged Microdata File." Paper presented at the 1979 Meetings of the American Statistical Association, Survey Research Methods Section, August 16. (S) Housni, El Arbi; Notzon, Samuel; and Fichet, Marie P-aniele (1978). "PGE/ERAD/ECD Matching Experience in Morocco," in Karol Krotki, ed., Developments in Dual System Estimation of Populatioit Size and Growth. The University of Alberta Press. (E) Jabine, Thomas B. (1976). "The Impact of New Legislation on Statistical and Research Uses of SSA Data." 1975 Proceedings of the American Statistical Association, Social Statistics Section, 221-30. (E) Kadane, Joseph B. (1975). "Statistical Problems of Merged Data Files." OTA Paper 6, Office of Tax Analysis, U.S. Treasury Department (December 12). (S) Kadane, Joseph B. (1978). "Some Statistical Problems in Merging Data Files," in 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (Kadane's reply to Sims also appears in that volume.) (S)

King, Jill A. (1977). "The Distributional Impact of Energy Polices: Development and Application of the Phase I Comprehensive Human

Resources Data System (Task 8 Final Report)." Project Report Series MPR/PR 77-13, Mathematica Policy Research, Inc. (June 30). (S)

Madigan, Francis C., and Wells, H.B. (1976). "Report on Matching Procedures of a Dual Record System in the Southern Philippines." Demography vol. 13 no. 1, pp. 381-395. August. (E)

Marks, Eli S.; Seltzer, William; and Kr(>tki, Karol J.

(1974). "Population Growth Estimation-A Handbook of Vital Statistics Measurement." The

Population Council, New York. (E)

Nathan, Gad (1978). "The Use of an Experimental Study of Reaching Decisions on Matching Rules," in Karol Krotki, ed., Developments in Dual System Estimation of Population Size and Growth. The University of Alberta Press. (E)

Neter, John; Maynes, E.S.; and Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors." Journal of the American Statistical Association, 60: 1005-1027. (E) Newcombe, Howard B., and Kennedy, James M. (I 962). "Record Linkage, Making Maximum Use of the Discriminating Power of Identifying Information."

Communication of the Association for Computing Machinery, 5: no. 11, 563-566. (Nov.) (E)

Newcombe, H.B.; Kennedy, J.M.; Axford, S.J.; and James, A.P. (1959). "Automatic Linkage of Vital Records." Science, 130, no. 3381, 954-9. (E)

Office of Federal Statistical Policy and Standards (1978 a). A Framework for Planning U.S. Fedral Statistics. U.S. Department of Commerce: 371-372.(S) Office of Federal Statistical Policy and Standards (1978 b). "Report on Statistical Disclosure and Disclosure-Avoidance Techniques." Statistical Policy Working Paper 2, U.S. Department of Commerce. (E,S)

Okner, Benjamin A. (1972). "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement (July) 1: 325-52. (Okner's reply to comments also appears in that issue.) (S)

Okner, Benjamin A. (1974). "Data Matching and Merging: An Overview." Annals of Economic and Social Measurement (April) 2: 347-

52. (E,S)

Pappas, Norma Gavin (1979). "Design for a "Selected Bibliography on the Matching of Person Merged Data Base for National Health Insurance Records from Different Sources." 1974 Proceed- and Medicaid Analysis." Mathematica Policy Research, Inc. (revised May). (S)

Peck, Jon K. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 347-8. (S)

Perkins, Walter M., and Jones, Charles D. (1966). "Matching for Census Coverage Checks." 1965 Proceedings of the American Statistical Association, Social Statistics Section, 122-41. (E) Radner, Daniel B. (1974). "The Statistical Matching of Microdata Sets: The Bureau of Economic Analysis 1964 Current Population Survey-Tax Model Match."

Ph. D. Dissertation, Department of Economics, Yale University. Microfilm. (S) Radner, Daniel B. (I 977). "Federal Income Taxes, Social Security Taxes, and the U.S. Distribution of Income, 1972." Paper presented at the 15th General Conference of the International Association for Research in Income and Wealth, University of York, England. August 19-25 (ORS Working Paper No. 7, Office of Research and Statistics, Social Security Administration, April 1979). (S) Radner, Daniel B. (1978). "Age and Family Income." Paper presented at the NBER Workshop on Policy Analysis with Social Security Research Files, Williamsburg, Virginia, March 15-17 (in Policy Analysis with Social Security Research Files, the proceedings of the workshop). (S) Radner, Daniel B. (1979). "The Development of Statistical Matching in Economics." 1978 Proceedings of the American Statistical Association, Social Statistics Section, 503-8. (S) Radner Daniel B., and Muller, Hans J. (1978). "Alternative Types of Record Matching: Costs and Benefits." 1977 Proceedings of the American Statistical Association, Social Statistics Section, 756-61. (E,S)

Ruggles, Nancy, and Ruggles, Richard (1974). "A Strategy for Merging and Matching Microdata Sets." Annals of Economic and Social Measurement (April) 2: 353-72. (S)

Ruggles, Nancy; Ruggles, Richard; and Wolff, Edward (1977). "Merging Microdata: Rationale, Practice and Testing." Annals of Economic and Social Measurement (Fall) 6: 429-44. (S)

Scheuren, Fritz and Oh, H. Lock (1976). "Fiddling Around with Nonmatches and Mismatches." 1975 Proceedings of the American Statistical Association, Social Statistics Section, 627-33. (E)

ings of the American Statistical Association, Social Statistics Section, 151-4 (Compiled by Fritz Scheuren and Wendy Alvey, Social Security Administration; 151 references). (E) Seltzer, William, and Adlakha, Arjun (1969). "On the Effect of Errors in the Application of the Chandrasekar-Deming Technique." (Reprinted as Laboratories for Population Statistics Reprint Series No. 14.) Chapel Hill, 1974. (E)

Sims, Christopher A. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 343-46. (Sims' "Rejoinder" also appears in that issue.) (S) Sims, Christopher A. (1974). "Comment." Annals of Economic and Social Measurement (April) 2: 395-8. (S) Sims, Christopher A. (1978). "Comments on Kadane's Work on Matching to Create Synthetic Data." in 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (S)

Smith, Martha E., and Newcombe, H.B. (1975).

"Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories." Methods of Information in Medicine (July) 14: 118-25. (E) Spiers, Emmett F., and Knott, Joseph J. (1970). "Computer Method to Process Missing Income and Work Experience Information in the Current Population Survey." 1969 Proceedings of the American Statistical Association, Social Statistics Section, 289-97. (S) Springs, Ricardo, and Beebout, Harold (1976). "The 1973 Merged SPACE/AFDC File: A Statistical Match of Data from the 1970 Decennial Census and the 1973 AFDC Survey." Mathematica Policy Research, Inc. (March 31). (S)

Steinberg, Joseph, and Pritzker, Leon (1967). "Some Experiences with and Reflections on Data Linkage in the United States." Bulletin of the International Statistical Institute, 42: 786-805. (E)

Tepping, Benjamin J. (1968). "A Model for Optimum Linkage of Records." Journal of the Amercan Statistical Association, 63: 1321-1332. (E)

Turner, J. Scott, and Gilliam, Gary E. (1975). "Reducing and Merging Microdata Files," OTA Paper 7, Office of Tax Analysis, U.S.

Treasury Department (October). (S)

United Nations Statistical Office (1978). "The Organization of Integrated Social Statistics." (Expert Group Meeting on Methods of Integration of Social and Demographic Statistics. 27-31 March) (February 27). (E,S)

U.S. Department of Agriculture, Statistical Reporting Service (1977). "Selection of a Sumame Coding Procedure for the SRS Record Linkage System." (B. T. Lynch and W. L. Arends). Paper presented at the April 6, 1977 meeting of the Washington Statistical Society. (E)

U.S. Department of Commerce, National Bureau of Standards (1977). "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers." NBS Special Publication 500-2. (E)

Wolff, Edward N. (1974). "The Goodness of Match." National Bureau of

Economic Research Working Paper No. 72 (December). (S) Wolff, Edward N. (1977). "Estimates of the 1969 Size Distribution of Household Wealth in the U.S. from a Synthetic Database." Paper presented at the Conference on Research in Income and Wealth, Williamsburg, Virginia, December. (S) Wyscarver, Roy A. (1978). "The Treasury Personal Individual Income Tax Simulation Model," OTA Paper 32, Office of Tax Analysis, U.S. Treasury Department (July). (S)

Reports Available In the Statistical Policy Working Paper Series

1. Report on Statistics for'Allocation of Funds
   GPO Stock Number 003-005-00178-6, price $2.40.

2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques GPO Stock Number 003-005-00177-8, price $2.50.

3. An Error Profile: Employment as Measured by the Current Population Survey

GPO Stock Number 003-005-00182-4, price $2.75.

4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (A limited number of free copies are available from OFSPS).

5. Report on Exact and Statistical Matching Techniques.

Copies of these working papers, as indicated, may be ordered from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Please use G.P.O. stock number when ordering.